A FUZZY LOGIC BASED FRAMEWORK FOR RELEVANT INFORMATION RETRIEVAL

THESIS

submitted in fulfilment of the requirement of the degree of

DOCTOR OF PHILOSOPHY

to

YMCA UNIVERSITY OF SCIENCE & TECHNOLOGY

by

MAMTA KATHURIA
Registration No: YMCAUST/PH53/2011

Under the Supervision of

Dr. C. K. NAGPAL
Professor

Dr. NEELAM DUHAN
Assistant Professor

Department of Computer Engineering
Faculty of Engineering and Technology
YMCA University of Science &Technology
Sector-6, Mathura Road, Faridabad, Haryana, INDIA
November 2018

DEDICATION

to

My Husband Mr. Naresh Kathuria

and

My beloved kids Parth Kathuria & Yuva Krishna

CANDIDATE'S DECLARATION

I hereby declare that this thesis entitled "**A FUZZY LOGIC BASED FRAMEWORK FOR RELEVANT INFORMATION RETRIEVAL" by MAMTA KATHURIA,** being submitted in fulfilmentof requirement for the award of Degree of Doctor of Philosophy in the Department of Computer Engineering under Faculty of Engineering and Technology of YMCA University of Science and Technology, Faridabad, during the academic year 2018-2019, is a bonafide record of my original work carried out under the guidance and supervision of **Dr. C. K. NAGPAL, PROFESSOR & Dr. NEELAM DUHAN, ASSISTANT PROFESSOR, DEPARTMENT OF COMPUTER ENGINEERING** and has not been presented elsewhere.

I further declare that the thesis does not contain any part of any work which has been submitted for the award of any degree either in this university or in any other university.

(MAMTA KATHURIA)

Registration No. YMCAUST/PH53/2011

CERTIFICATE

This is to certify that this thesis entitled **"A FUZZY LOGIC BASED FRAMEWORK FOR RELEVANT INFORMATION RETRIEVAL"** by **MAMTA KATHURIA** submitted in fulfilment of the requirement for the award of Degree of Doctor of Philosophy in **DEPARTMENT OF COMPUTER ENGINEERING**, under Faculty of Engineering and Technology of YMCA University of Science and Technology, Faridabad, during the academic year 2018-2019, is a bonafide record of work carried out under my guidance and supervision.

I further declare that to the best of my knowledge, the thesis does not contain any part of any work which has been submitted for the award of any degree either in this university or in any other university.

Dr. C. K. Nagpal
PROFESSOR
Department of Computer Engineering
Faculty of Engineering and Technology
YMCAUST, Faridabad

Dr. Neelam Duhan
ASSISTANT PROFESSOR
Department of Computer Engineering
Faculty of Engineering and Technology
YMCAUST, Faridabad

Date:

ACKNOWLEDGEMENT

It gives me immense pleasure to acknowledge the people whose priceless contributions helped me reach here. Without their support, this piece of academic work, in the form of my doctoral thesis, would have just been reduced to mere ink on paper. First and foremost, my deepest and heartfelt thanks to Dr. C. K. Nagpal who has been not just a guide par excellence but also an inspiring teacher, Nagpal sir not only created a solid foundation in my mind , but also nurtured it over the years till this day through his tireless efforts.

Sir, you have shown me the right path always, lit the lamp of clarity whenever I was in doubt, corrected me when I was wrong, encouraged me when I was right, criticized me when I became over-confident and motivated me whenever I would lose hope.

I express my deep sense of gratitude to Dr. Neelam Duhan, for her valuable guidance and help without which it was not possible for me to complete my work. Mam, you believed in my abilities and gave unconditional support to help me achieve excellence throughout my research work culminating in this doctoral thesis.

It is my firm belief that no success in life can ever be achieved without the well wishes and support of one's family. And when I look back from where I have reached today, my belief only becomes stronger. First and foremost, my very special remembrance and gratitude towards my father, Sh. Subhash Mehta, for believing in me and giving me the best lesson of life. My father inspired me to do better with every passing day. I am sure no one today would be as proud as him seeing me complete my doctoral thesis. I also heartily thank my dearest mother, Mrs. Kamlesh Rani, Father- in- law Sh. Roop Chand and mother- in- law Mrs. Kaushalya, Brothers Prayag and Anand, Sister Neha for their unwavering support. And even though my sons Parth Kathuria and Yuva Krishna are too young to understand these words, thanks to them for bringing smiles on my face by coming into our lives.

People say that behind every successful man there is a woman, but in my case it has been the other way round. No matter what I write, no words can even suffice my gratefulness for the man of my life, my best friend, my biggest support, a never-ending source of inspiration and righteousness, my soul mate and life partner Mr.

Naresh Kathuria. He gave me hope, courage and the reason to be what I am today, my true self and he always said, “Be the best at what you do”. Today, I know I am making him proud.

I would like to express my sincere gratitude to chairpersons Dr. Atul Mishra and Dr. Komal Kumar Bhatia for their valuable advice, support and information on different aspects.

I would also like to thank to my colleagues Dr. Shilpa and Ms. Amita, my dearest friend Ms. Deepshika, my fellow students Anuradha, Riya and Kalyan Mudireddy for all the help, discussion and great company.

And finally, thinking that one entity without whose grace, all the above wonderful people wouldn’t have come in to my life and without whose blessings, I wouldn’t have been where I am. Thank you almighty, the ever knowing, omnipresent and ever-forgiving God for being kind and watching over me all these years and I know that you will continue to do so forever.

(MAMTA KATHURIA)

Registration No. YMCAUST/PH53/2011

ABSTRACT

The performance of search engines in today's scenario is quite impressive yet there has been the ever felt need for novel mechanisms for executing/realizing users' expectations seeking the rich set of relevant results for the query submitted by them. The massive size, continuous update of the information, heterogeneity on the basis of various factors like linguistics, geographical location, cultures and other parameters make the task of information retrieval quite complex and challenging.

Most of the web search engines are based upon the query text which is a very short piece of natural language expression. The ambiguity in the natural language, conceptual references, entity references and synonyms thereof add the complexity to the matter. To ensure the rich and relevant results in response to submitted query, which is a very short piece of text, it is desirable that search engine must be able to rephrase/expand the query through its multiple versions with each version containing the quality synonym for the entity and the attribute references. The search engine must also be in a position to translate a given concept to its appropriate set of instances using worldly knowledge.

The work carried out in this thesis involves generating synonyms of the attribute word present in the query through the identification of their contexts. The synonyms so identified have the global implications based upon the billions of pages instead of individual perception. The methodology used not only creates the synonyms but also provides an index to assess the similarity in meaning. The index was normalized and a fuzzy rule base was created for the purpose of usage in automation process.

The work also involves the creation of set of synonyms for the entity references. In contrast to the attribute words, there doesn't exist any lexical reference to find the seed for entity synonym. The proposed work is based upon the web data and web logs containing rich set of information. This work is an improvement over the existing works both in terms of quality and quantity. This has been proved through the generation of an index that measures the similarity between two entity synonyms.

The work was further extended to realization of the conceptual references to their appropriate instances through the use of latest available worldly knowledge

repository, PROBASE, in the light of user centric data such as browsing history, geographical location, IP address etc.

The work explores the alternative representations for *attributes, entity references* and *concepts* components of a query. *Keywords* which carry the essence of the query were not considered for alternative representation. The alternative representations proposed in this work shall help the search engines in meaningful rephrasing of the query to ensure rich and relevant information as sought by the user.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation	Details or Expanded Form
WWW	World Wide Web
N.A	Not Available
FRB	Fuzzy Rule Base
STS	Semantic Textual Similarity
SVR	Support Vector Regression
NGD	Normalized Google Distance
CODC	Co-Occurrence based Double-Checking
IPC	Interacting Page Count
ICR	Intersecting Click Ratio
SERPs	Search Engine Result Pages
FCS	Fuzzy Control System
URL	Universal Resource Locators
SPU	Sub Parent URL
RDF	Resource Description Framework
OWL	Web Ontology Language
SPARQL	Simple Protocol and RDF (Resource
OMCS	Open Mind Common Sense
NELL	never-ending language learning
CIM	Concept Identification Module
CRM	Concept Resolution Module
RPM	Result Processing Module
CERF	Concept Entity Relationship File
IP	Internet Protocol

CHAPTER I
INTRODUCTION

1.1 SEARCH ENGINES & INFORMATION RETRIEVAL

***The World Wide Web* (*WWW*)** is a gigantic repository that keeps information related to almost every domain of knowledge accessible everywhere on anytime basis. The massive size, continuous update of the information, heterogeneity on the basis of various factors like linguistics, geographical location, cultures and other parameters make the task of information retrieval quite complex and challenging.

Though the performance of search engines in today's scenario is quite impressive yet there has been the ever felt need for novel mechanisms for accomplishing expectations of the users who are seeking the rich set of relevant results for their submitted queries.

The basic reasons for the inability of the search engines to provide the relevant results which are not up to expected levels are as follows:

- Query is a very short piece of text in natural language and successful retrieval is very much dependent of the intent of the user behind the query.
- Natural language is ambiguous and affects the relevance/quality of search results returned by the search engine.
- Users may use slang terms which are not as such part of the language.
- The reference in the query may be conceptual requiring proper instantiation.
- The reference in the query may refer to an entity recognizable by different names.
- Current Lexical resources are unable to cover the heterogeneity of the web.
- Web is continuously updating.

All these issues need to be addressed for getting the rich and relevant information from the web. The literature contains a lot of work in this regard, the study of which motivated us to carry out the work proposed in this thesis.

1.2 PROBLEM IDENTIFICATION

To understand the ongoing work being carried out to overcome the above mentioned problems, nearly 100 research papers, as listed in the reference section and discussed in brief in Chapter II of this thesis, were studied. It was felt that there is an ample opportunity to carry out further research work to ensure the rich and relevant information by working on the various components of the query. The literature survey has shown that the basic constituents of a query can be classified into four categories: *Keywords, Attributes, Entities and Concepts*.

(i) *Keywords* are non-trivial words which carry the essence of the query. The keywords make the query meaningful and are the major guiding factors for relevant information retrieval to be carried out by the search engines.

(ii) *Entities* are persons/places/objects referred in the query which have distinct and independent existence. Different users may refer to the same entity in different manners. For example, the newspaper *The Times of India* may also be referred to as *TOI*. A search engine must be able to handle these multiple versions of the references used in the query. These multiple references have been referred to as *entity synonyms* in the work [1, 2]. Creation of appropriate set of entity synonyms for a given entity is also a major requirement for relevant and rich information retrieval. Various contributions in this field are available in [1-9].

(iii) *Attributes* are the words which define the features/ characteristics of entities and keywords used in the query. To enrich the search process, a web search engine may create multiple versions of the same query by using the appropriate set of synonyms of the attributes used in the query and create an index to access the quality of the synonym generated. Various contributions in this field are available in [10-18].

(iv) A *Concept* in the query is a word which refers to a broad category of objects in generic manner. For example, in the query "*good actors in India" Good actors* is a concept. A concept referred in query has to be translated to its closest set of instance(s). Handling of the concept is the most challenging task for the search engine as its resolution requires the understanding and usage of worldly knowledge. The instantiation of a

concept can vary depending upon the local & global contexts. The hardest part of the query expansion is to find the appropriate instantiation for the concept used in the query as per the requirement of the user. Various contributions in this field are available in [20-32].

After going through the literature, following inferences were drawn:

- Keywords are the essential part of the query and should not be disturbed/modified/altered.
- Lexical resources are unable to provide the requisite set of synonyms for the words used in the query owing to the widespread and heterogeneous nature of the web. So, there is a need to find out global mechanisms for creating the set of synonyms which truly cover the heterogeneity of the web.
- Alternative references to entities (also known as entity synonyms) are not at all supported by the lexical resources. The only way one can find out the entity synonyms is through web exploration and analysis of web logs.
- Conceptual references need to be translated into their worthy instances which are quite a challenging task as it requires worldly knowledge.

After exploring all this literature, we were in a position to set the objectives of the proposed work.

1.3 PROBLEM DEFINITION

To ensure relevant web search through query rephrasing or expansion using

- rich set of identified synonyms for the entities and the attributes used in the query
- appropriate instances for the concepts present in the query

and to devise novel mechanisms for the purpose.

1.4 OBJECTIVES

Following objectives were set for the proposed work:

a) To devise a mechanism to search synonyms of an attribute word of the query using huge document repositories.
b) To devise a mechanism to search rich set of entity synonyms for an entity using static and dynamic web.
c) To design an index to assess the quality of synonyms as two synonyms of the same word can't have same intensity.
d) To devise a mechanism to translate a concept to its intended instances using worldly knowledge source.
e) To devise a mechanism for automated usage of identified set of synonyms to be utilized by the machine.

1.5 ACCOMPLISHMENTS

Following accomplishments were made during this work:

a) A mechanism to search synonyms of an attribute word on the basis of the context identification using multiple corpora was proposed and implemented. The method is quite an improvement over the existing methods which use page count and snippets.
b) A mechanism to generate rich set of entity synonyms for an entity using query log and anchor text was proposed and implemented.
c) For both of the above mechanisms, an index was created to assess the quality of synonyms. The index was fuzzified and a Fuzzy Rule Base (FRB) was created for automated deployment of synonyms for various purposes.
d) A mechanism to translate a concept to its intended instances was proposed and implemented using PROBASE (the largest available worldly knowledge source) [33].

1.6 ORGANIZATION OF THE THESIS

The thesis has been organized as follows:

Chapter II: Literature Survey: This chapter contains a discussion on the available work related to search engine evolution, semantic similarity between words, entities and concept based web search. Based on the literature survey on each topic, the

problems and challenges have been identified and discussed in brief. These problems and challenges form the basis for the work carried out.

Chapter III: Synonym Resolution for Attribute Component in Query: This chapter talks about the proposed semantic similarity technique and its implementation for attributes component present in the query. To assess extent of similarity between the synonyms under consideration and the candidate word, list of their contexts have been taken into consideration. The work makes use of various corpora for extracting contexts.

Chapter IV: Dynamic Entity Resolution: This chapter covers the detailed discussion on the proposed and implemented work to generate a rich set of entity synonyms for the commonly used entities using web data, web log and anchor text. An index has also been created to assess the quality of the created synonym. Obtained results have been compared with the existing techniques.

Chapter V: Concept Resolution for Focused and Enriched Web Information Retrieval: This chapter proposes and implements an algorithm for concept resolution using PROBASE, a huge taxonomy on worldly knowledge created by Microsoft, in combination with users' statistics resulting in focused and enriched outcomes. The results so obtained have been compared with the outputs of existing search engines such as Google, Bing and Yahoo.

Chapter VI: Conclusion and Future Scope: This chapter concludes the work and provides a description of potential future work in the area under consideration.

We move to next chapter that presents existing literature related to problem taken up in this work.

CHAPTER II

LITERATURE SURVEY

2.1 INTRODUCTION

In this chapter, the study of existing work carried out by various researchers to make the information retrieval precise as per the requirement of the web searcher is taken up. The process started with the study of the search engines evolution process from their infancy to the current scenario. For this purpose, many websites and research papers (more than 100) were referred to. The outcome of the study has been shown in Table 2.1.

Table 2.1 Web Search Engines Evolution Process

Sr. No.	Year/Search Engine Name	Key Developer / Developed at or Owner	Features and Innovations	Current Active status/ Alexa Rank
1.	1990 Archie[34]	Alan Emtage, Peter J. Deutsch, Bill McGill University, Montreal	1. FTP Server based sharing of files 2. crawling concept 3. Script-based data gatherer 4. Regular Expression based matching retrieval of files for user query	Not Active Alexa N.A
2.	1992 Veronica & Jughead [35]	Fred Barrie, Rhett Jones University of Naveda System Computing Services group	1. Menu Driven approach 2. Ability to search plain text files 3. Keyword based search in 4. Its own designed Gopher Index System	Not Active Alexa N.A
3.	1993 W3 Catalog [36,37]	Oscar Nierstrasz University of Geneva	1. Purely textual browser 2. Integration of manually maintained catalogue. 3. Dynamic querying	Not Active Alexa N.A
4.	1993 JumpStation [38]	Jonathon Fletcher University of Stirling	1. Combines crawling, searching and indexing 2. Lays the foundation for current form of search engines 3. Unable to grow because of linear search drawback	Not Active Alexa N.A
5.	1993 WWW Wanderer [39]	Matthew Gray Massachusetts Institute of Technology	1. Introduces web robots to crawl the web 2. Track the web's growth, Indexed titles and URLs 3. Did not facilitate web search, major goal to measure web size 4. Perl based web crawler	Not Active Redirected to Yahoo, Alexa N.A
6.	1993 Aliweb [40]	Martijn Koster United Kingdom	1. Devoid of crawling mechanism 2. Website administrator had to register with Aliweb to get their services listed & indexed 3. Capability to perform Archie Like Indexing for the web	Active(www. aliweb.com Alexa N.A

7.	1994 Web Crawler [41]	Brian Pinkerton	1. Lays the foundation for Content Based Search 2. Use of Boolean operators in user query 3. User Friendly Interface	Active, Aggregator, https://www webcrawler.com/ 674
8.	1994 Meta Crawler [42]	Erik Selberg, Oren Etzioni Blucora Inc.	1. Introduced the concept of meta search wherein search results of major search engines are combined to widen the search results. 2. Does not have its own search index	Active, Aggregator, http://www.metacrawler.co.uk/ 8688
9.	1993 Myweb Search [43]	IAC	1. Search tool compatible with Internet Explorer (4.x or above) and Netscape 4.x. 2. It is a spyware and search toolbar program 3. Displays algorithmic search results from Google, Ask.com, Yahoo and LookSmart, along with sponsored listings, primarily from Google. 4. Easy to add/remove additional software products to the Toolbar. 5. Free to use	Active but powered by google http://home.mywebsearch.com/index.jhtml 405
10.	1994 Lycos [44]	Mauldin Micheal L. Canegie Mellon Univ. , Pittsburg	1. Prefix matching and word Proximity 2. Keyword, search on image or sound files 3. Focuses more on directory	http://www.lycos.com/Search/ 9041
11.	1994 Inktomi [45]	Eric Brewer University of California	1. First major search engine to launch a paid inclusion service 2. Handles thousands of search queries by distributing among many servers	Not Active, Aquired by Yahoo, Alexa N.A
12.	1994 Infoseek [46]	Steve Kirsch Infoseek Corporation	1. Provided subject oriented search 2. Allowed real-time submission of the page	Not Active Alexa N.A
13.	1995 Excite [47]	Joe Kraus, Graha spencer Garage in Silicon velley	1. Both concept & keyword based search 2. Large & up-to-date index 3. Excellent summaries 4. Fast, flexible, reliable searching 5. Idea of statistical analysis of word relationship for efficient search	Active, Now an internet Portal http://www.excite.com/ 7951
14.	1995 AltaVista [48]	Louis Monier, Michael Burrows Digital Equipment Corporation's	1. Fast Multithreaded crawler & Back-end search 2. Keyword based simple or advanced search 3. Multilingual search capabilities 4. Periodic Re-indexing of sites 5. High bandwidth 6. Allow natural language query 7. Inbound link checking	Not Active, Shutdown in 2013, redirected to Yahoo, 565211
15.	1995 Yahoo [49, 50]	David Filo, Jerry Yang Yahoo Corporation	1. Keyword based search 2. Web directory organized in hierarchy 3. Separate searches for images, news stories, video, maps, shopping 4. Supports full Boolean searching 5. Support Wild Card Word in Phrase	2nd largest Active SE https://in.yahoo.com/ 4

16.	1995 AOL [51]	Bill von Meister Control Video Corporation	1. Started as Internet 2. Messenger Service 3. Subscriber based service 4. Movie & Game portal	Not Active http://www.aol.in/, Alexa N.A
17.	1995 MSN [52]	Microsoft Microsoft ltd.	1. Large and unique database 2. Boolean searching 3. Cached copies of Web pages including date cached 4. Automatic local search options. 5. Neural n/w added features	Active as Bing http://www.msn.com/en-in/, 48
18.	1996 DogPile [53,54]	Aaron Flin Blucora Inc.	1. Meta Search engine 2. Has its own search Index 3. Searched multiple engines, filtered for duplicates and then presented the results to the user 4. Special provisions for Stock quotes, weather forecast, yellow pages etc.	Active, Aggregator http://www.dogpile.com/ 3084
19.	1996 InfoSpace [55]	Naveen Jain Infospace Inc.	1. Meta Search Engine 2. Selects results from the leading search engines and then aggregates, filters and prioritizes the results to provide more comprehensive results 3. Instant messenger service	Active http://infospace.com/ 2110
20.	1996 Hotbot [56,57]	Wired Magazine Inktomi Corporation	1. Extensive use of cookie technology to store personal search preference information 2. Ability to search within search results 3. Frequent updation of Database Use of parallel processing	Active http://www.hotbot.com/ 100902
21.	1996 WOW [58]	Jeniffer Thompson Compu Serve	1. First internet service to be offered with a monthly "unlimited" rate 2. Brightly colored 3. Seemingly hand-drawn pages. 4. Find all of the breaking news articles, top videos and trending topics that matter to you. 5. Effective advertising 6. Highly communicative design 7. Budget friendly media services 8. Creative concept development	Active http://www.wow.com/ 767
22.	1996 Ask [59,60]	DavidWarthen, Garrett Gruener IAC/ InterActive Corporation	1. Natural language-basedSearch 2. Both concept & keyword based search 3. Allows to enter query in the form of sentence for humanize the online experience 4. Question answering system	Active http://www.ask.com/ 28
23.	1997 Daum [61]	Daumkakao Daum Corparation	1. A popular search engine in Korea 2. Besides internet search provides facilities for E-mail, Chat, Shopping etc.	Active www.daum.net/ 140
24.	1997 Overture [62]	Bill Gross Yahoo	1. Paid search inspired from commercial telephone directory 2. Secured, pay-per-placement directory service	Not Active Alexa N.A

25.	1997 Yandex [63]	Taylor Nelson Sofres San Francisco Bay Area	1. Full-text search with Russian morphology support 2. Encrypted search 3. Multilingual	Active https://www.yandex.com/ 20
26.	1998 Google [64,65]	Sergey Brin, Lawrence Stanford University, Stanford	1. Keyword based search 2. Page Rank algorithm 3. Semantic search 4. Free, Fast and easy to search 5. No programming or database skills required	Active as most popular SE https://www.google.co.in/, 1
27.	1999 AlltheWeb [66]	Tor Egge Norwegian Univ. of Sci. & Tech.	1. Faster Database 2. Advanced search features 3. Sleek interface 4. FAST's enterprise search engine 5. search clustering 6. completely customizable look	Not Active (URL redirected to Yahoo), Alexa N.A
28.	2000 Teoma [67]	Apostolos Gerasoulis Rutgers Univ. computer lab	1. Provide knowledge search 2. Provide subject specific popularity 3. Clustering Techniques to 4. Determine Site Popularity 5. Unique Link popularity	Not Active , Redirected to Ask.com, Alexa N.A
29.	2000 Baidu [68]	Robin Li Beijing China	1. largest internet user population 2. pay per click marketing platform 3. China's Google	Active http://www.baidu.com/ 5
30.	2007 LiveSearch [69]	Satya Nadella Microsoft	1. Uses a drag-and-drop interface that's really simple to pick up 2. The new search engine used search tabs that include Web, news, images, music and desktop	Active as Bing, Launched as rebranded MSN search https://www.live.com/, Alexa N.A
31.	2008 DuckDuckGo [70]	Gebriel Weinberg DuckDuckGo Inc.	1. Offers real privacy or protecting searchers' privacy and avoiding the filter bubble of personalized search results 2. Smarter search, and stories that user likes 3. Not profiling its users and by deliberately showing all users the same search results for a given search term 4. Emphasizes on getting information from the best sources rather than the most sources	Active https://duckduckgo.com/ 506
32.	2008 Aardvark [71]	Max Ventilla, Nathan Stoll The Mechanical Zoo, A San Francisco based startup	1. Use Social n/w facilitated a live chat or email conversation with one or more topic experts 2. Social search Engine 3. Aadvark Ranking Algorithm	Not Active Alexa N.A
33.	2009 Bing [72]	Steve Billmer Microsoft	1. Keyword based search 2. Index updated on weakly or daily basis 3. Advertised as a decision engine 4. Social integrations are stronger 5. Direct information in the area of finance & sports	Active https://www.bing.com/ 24

34.	2009 Caffeine [73]	Matt Cutts Google	1. New web indexing system 2. Near-real-time integration of indexing and ranking 3. Allows easier annotation of the information stored with documents 4. Provide 50% fresher result 5. Find links to Relevant content much sooner 6. Update search index on a continuous basis, globally. 7. Caffeine processes hundreds of thousands of pages in parallel. 8. Nearly 100 million gigabytes of storage in one database	Active http://googleblog.blogspot.in/ 2010/06/our-new-search-index-caffeine.html, Alexa N.A
35.	2010 Google Instant [74]	Marissa Mayer & Matt Cutts	1. Search-before-you-type 2. Predicts the users whole query 3. Faster Searches, Smarter Prediction, Instant Result 4. User Experience 5. Provide Auto complete Suggestion	Active, Alexa N.A
36.	2010 Blekko [75]	Rich Skrenta Blekko Inc.	1. Uses slash tags toallow people to search in more targeted categories 2. Spam Reduction 3. Provides better search results than those offered by Google Search, by offering results culled from a set of billion trusted websites and excluding material from such sites as content farms. 4. Dynamic interface graph algorithm 5. Blekko offers a web search engine and social news platform that provides users with curated links for the entered search criteria. 6. Provides downloadable search bar which was later acquired by IBM	Active, Aquired by IBM(www.blekko.com) 4518
37.	2013 Contenko [76]	Tomas Meskauskas Amerow LLC	1. Deceptive Internet Search, promoted using various browsers hijackers 2. Provides Innovative means for browsing the internet 3. Its Startup page doesn't contain any links to privacy terms or terms of use	Active http://www.contenko.com/ 4505
38.	2013 Alhea [77]	Manuel Barrios Amazon Technologies Inc.	1. Offers a single source to search the Web, images, audio, video, news from Google, Yahoo!, and many more search engines. 2. Alhea.com compiles results from many of the Web's major search properties, delivering	Active http://www.alhea.com/ 11225
9.	2011 GooglePanda [78]	Navneet Panda and Vladimir Ofitserov Google	1. Focuses on eliminating sites that didn't have enough quality content and were more geared at moneymaking than providing useful content. 2. Provides new Google's search results ranking algorithm 3. Quality Search results	Active http://www.google-panda.com/, Alexa N.A

40.	2012 GooglePenguin [78]	Matt Cutts Google	1. Web spam update 2. Goal of concentrating on webspam 3. Search Algorithm update 4. Protect your site from bad links .	Active Alexa N.A
41.	2013 Google HummingBird [78]	Gianluca Fiore Lli	1. A core algorithm update may enable more semantic search and more effective use of the Knowledge Graph in the future, Hummingbird is about synonyms but also about context Google 2. Hummingbird is designed to apply the meaning technology to billions of pages from across the web, in addition to 3. Knowledge Graph facts, which may bring back better results 4. Search Algorithm update 5. Understand the intent of the user	Active Alexa N.A
42.	2015 SciNet [79]	TuukkaRuotsalo, KumaripabaAthukorala, DorotaGłowacka, KseniaKonyushkova, AnttiOulasvirta,SamulimKaipiainen, SamuelKaski, GiulioJacucci Helsinki Institute for Information Technology HIIT, Finland	1. Reinforcement Learning 2. Auto-suggestion for specific topic & document 3. Interactive approach 4. A new search engine that outperforms current ones and helps people search more efficiently. 5. SciNet displays a range of keywords and topics in a topic radar	Active 2159988
43.	2016 Clusty [80]	Carnegie Mellon University researchers	1. Clusty is a clustering engine that groups similar items together – organizing search results into folders. 2. It combines the power of clustering with meta-search 3. It provides a productive and flexible search experience. 4. It produces organic web results 5. It enables searching of shopping information, yellow pages data, news, blog posts and images. 6. The competition has shifted from crawling the web and returning search results, to adding value to the information that has been retrieved. Clusty has a few advantages over Google: a. You don't have to come up with your own categories or subjects in order to narrow, or refine, the search. b. You don't have to rely on Google's perceived emphasis on links. c. You don't have to guess the keyword, to get to that perfect page you need. Navigate the clusters and sub-clusters, just as you would use eBay, to find that one specific treasure you've been hunting for.	Active, https://searchenginewatch.com/tag/clusty/ 117,450

44.	2005 Lexxe [81]	Hong Liang Qiao, Australia	1. Internet search engine that applies Natural Language Processing in its semantic search technology. 2. Offers Linguistic Search 3. The Questions and answer search engine uses linguistics to answer the questions that are posed as queries.	Active, https://in.linkedin.com/company/lexxe-search Alexa N.A

2.2 DEFINING THE SCOPE OF THE PROBLEM

The task to be accomplished in this work was to design an efficient framework for relevant information retrieval from the World Wide Web (WWW). The basis of the information retrieval is a query which is a short piece of text and the search engine has to analyze this short piece of text to ensure unambiguous, precise and rich information as per the requirement of the user. Since query is a very short piece of text, selecting an appropriate set of web pages is an uphill task. This led to the need for query recommendation in terms of Expansion/ Rephrasing/ Reformulation of the query that helps search engines in creating multiple versions of the input query, using polysemy and synonymy, to accomplish a rich and relevant information retrieval. These techniques use Query Expansion [82, 83], Association Rules [84], Query-Flow Graph [85] and Query Clustering [86].

To understand the crux of the matter, many papers were studied and we got concrete information in a paper [27,87] which conveyed that contents of the query text can be classified into four major types: *Keywords, Entities, Attributes* and *Concepts* with description as follow:

(i) *Keywords* are non-trivial words which carry the essence of the query. The keywords make the query meaningful and are the major guiding factors for relevant information retrieval to be carried out by the search engines.

(ii) *Entities* are persons/places/objects referred in the query which have distinct and independent existence. Different users may refer to the same entity in different manner. For example, the newspaper *The Times of India* may also be referred to as *TOI*. A search engine must be able to handle these multiple versions of the references used in the query. These multiple references have been referred to as *entity synonyms* in the literature. Creation of appropriate set of entity synonyms for a given entity is also a

major requirement for relevant and rich information retrieval. After studying the various contributions in this field we proposed a new strategy in this domain as discussed in Chapter IV of this thesis.

(iii) *Attributes* are the words which define the features/ characteristics of entities and keywords used in the query. To enrich the search process, a web search engine may create multiple versions of the same query by using the appropriate set of synonyms of the attributes used in the query. This necessitates the creation of the appropriate set of synonyms for a given attribute and to create an index to access the quality of the synonym generated. After studying the various contributions in this field we proposed a new approach in this domain as discussed in Chapter III of this thesis.

(iv) A *Concept* in the query is a word which refers to a broad category of objects in generic manner. For example, in the query "*good actors in India", Good actors* is a concept. A concept referred in query has to be translated to its closest set of instance(s). Handling of the concept is the most challenging task for the search engine as its resolution requires the understanding and usage of worldly knowledge. The instantiation of a concept can vary depending upon the local and global contexts. The hardest part of the query expansion is to find the appropriate instantiation for the concept used in the query as per the requirement of the user. After studying the various contributions in this field we proposed a new strategy in this domain as discussed in Chapter V of this thesis.

Our work in this thesis concentrates upon the query expansion and resolution process by using appropriate set of synonyms for attributes & entities and appropriate set of instances for concept resolution. Since different synonyms created for an input entity/attribute may not have same extent of similarity with the input entity/attribute, therefore, an index for assessing this extent of similarity was created which was fuzzified after normalization process. This fuzzification helped us in creating a Fuzzy Rule Base (FRB) which can be used for automated usage in the web search process. All this work leads to the creation of a framework for rich and relevant information retrieval from the web.

2.3 LITERATURE SURVEY FOR ATTRIBUTE SEMANTIC SIMILARITY

The inadequacy of the manually created lexical resources has been long felt due to their inability to cater various diversified domains of knowledge area such as engineering, medical, music and finance etc. Moreover, mostly their updation is also based upon the perception of a few people without much usage of statistics and engineering. With the growth of web over multifaceted heterogeneous environment spread over variety of domains, people, nationalities, dialects etc., the realization has become more and more prominent wherein these resources are unable to provide the requisite support to the web search engines for the purpose of meaningful query expansion. Thus, with the exponential growth of web, the need for meaningful query expansion has gained more and more prominence. This issue can only be addressed through proper semantic resolution which in turn requires knowledge about exact estimate of the semantic similarity between the terms. The growth of the web has pushed the research in the area of precise semantic similarity measurement to help target web users in getting the accurate results through proper semantic interpretation of their queries.

Major contributions in this field are as follows:

(a) Pilehvar, Jurgens, & Navigli (2013)[88] presented a unified approach for semantic similarity finding that operates at multiple levels from comparing word senses to comparing text documents. The method leverages a common probabilistic representation over word senses in order to compare different types of linguistic data. This unified representation shows state-of-the-art performance on three tasks: semantic textual similarity, word semantic textual similarity and word sense coarsening.

(b) Severyn, Nicosia, & Mchitti (2013)[89] have taken up the task of (STS) using a large number of pair-wise similarity features. Their model includes: encoding of input texts into relational syntactic structures, use of tree kernels to handle feature engineering, combining both structural and feature vector representations in a single scoring model using Support Vector Regression (SVR). The contribution of the work is quite significant in the area of semantic textual similarity.

(c) Specia, Jauhar & Mihalcea (2012)[90] provided a mechanism for Lexical Simplification which involves the replacement of words and phrases through their simpler variants using complexity analysis, substitute lookup and context-based ranking.

(d) Jurgens, Mohammad, Turney & Holyoak (2012)[91]have focussed on relational similarity such as entity:sound (dog:bark, cat:meow), cause:effect (virus:flu) etc. for the purpose of semantic similarity resolution.

Above mentioned approaches were quite academic and based upon classical concepts and statistics. The subsequent approaches in semantic similarity resolution involve the use of web contents to find the extent of semantic similarity. The inherent advantage of such an approach is its ability to take into consideration the heterogeneity of the web making them useful for the information retrieval from the web.

Initial efforts to make a decision about semantic similarity based upon web content were based upon page count. In these cases page count of the word pair (say P and Q) is taken into consideration for computation of semantic similarity. Here, H(P) and H(Q) indicate the page counts for word P and Q respectively and H(P $\cap$ Q) denotes the page count for conjunctive query 'P AND Q'. Based upon the page count, various indices have been proposed, which are as follows:

$$WebJaccard(P,Q) = \begin{cases} 0, & if\ H(P \cap Q) \leq c\ , \\ \frac{H(P \cap Q)}{H(P)+H(Q)-(H(P \cap Q)}, & otherwise \end{cases} \quad (2.1)$$

This coefficient is set to zero if page count H(P$\cap$Q) for the conjunctive query is less than a threshold c.

$$WebOverlap(P,Q) = \begin{cases} 0, & if\ H(P \cap Q) \leq c\ , \\ \frac{H(P \cap Q)}{\min\ (H(P),H(Q))}, & otherwise \end{cases} \quad (2.2)$$

$$WebDice(P,Q) = \begin{cases} 0, & if\ H(P \cap Q) \leq c\ , \\ \frac{2H(P \cap Q)}{H(P)+\ H(Q)}, & otherwise \end{cases} \quad (2.3)$$

The fourth measure WebPMI (Pointwise Mutual Information) reflects the independence between two probabilistic events and is defined as:

$$WebPMI(P,Q) = \begin{cases} 0, & if\ H(P \cap Q) \leq c\ , \\ log_2\left(\frac{H(P \cap Q)/N}{H(P)/N \cdot H(Q)/N}\right), & otherwise \end{cases} \quad (2.4)$$

where N is the number of documents being indexed by the search engine. Probabilities are assessed based on maximum likelihood principle.

With the advancement in search engine technology, new empirical methods were defined in the light of search results.

Cilibrasi & Vitanyi (2007)[92] defined a semantic similarity measure based called Normalized Google Distance (NGD) based upon page count provided by the search engine. The NGD measure can be defined as given in (2.5).

$$NGD(P,Q) = \frac{\max\{\log H(P), \log H(Q)\} - \log H(P,Q)}{\log N - \min\{\log H(P), \log H(Q)\}} \quad (2.5)$$

All these methods consider the co-occurrence of words without taking into account the context of the words in which they occur on the page.

2.3.1 Drawbacks of Page Count Based Measures

The major drawback of all these methods is that in many cases, two words may appear together on some pages even if they are not related, owing to distance between them in the page. Due to this reason, the page count based methods were considered to be quite crude and needed refinement through the augmentation of proper context. Thus, new methods were suggested based upon the text snippets where co-occurrence is taken into account only if it exists within a text snippet making it more relevant. These types of methods are outlined below.

(a) A double-checking model using text snippets was suggested by Chen, Lin & Wei (2006)[93] for measuring semantic similarity between two words. For computing the similarity between words P and Q, occurrences of word P are counted in the snippets for word Q and similarly the occurrences of word Q are counted in the snippets for word P. Then, the corresponding values are used to calculate the similarity between P and Q by using Co-occurrence based Double-Checking (CODC) measure as defined below:

$$CODC(P,Q) = \begin{cases} 0, & if\ f(P@Q) = 0, \\ exp\left(log\left[\frac{f(P@Q)}{H(P)} \times \frac{f(Q@P)}{H(Q)}\right]^{\alpha}\right), & otherwise \end{cases} \quad (2.6)$$

where f(P@Q) is the number of occurrences of P in the top ranking snippets of Google corresponding to query Q, H(P) denotes the page count for query P, and α is a constant which was experimentally set to the value 0.15 by the authors. The drawback of the method is its dependency on the ranking method adopted by a search engine in use. It happens because a search engine considers publication date and link structure along with semantic similarity while ranking the search results corresponding to the user query.

(b) Bollegala, Matsuo, and Ishizuka (2011)[10] proposed an empirical method to measure semantic similarity between two words based on page counts and text snippets retrieved from a search engine in response to the said words. They identified several semantic relations between given words by the using pattern extraction and pattern clustering algorithm. They also used support vector machine based supervised learning model for deciding the appropriate boundaries of classification.

(c) The work was further extended by Rada, Mili, Bichnell & Blettner (1989)[94] taking into consideration the taxonomy of words. According to them, shorter the path in taxonomy, more similar the words are.

(d) Resnik (1995)[95] proposed semantic similarity measurement between two concepts using Brown Corpus [96] and WordNet (2005)[97]. The proposed method finds the similarity between the concepts C1 and C2 on the basis of a third concept C that subsumes both the concepts and is the biggest of all such possible contents.

2.3.2 Limitations of Existing Word Similarity Methods

Above mentioned methods have the following drawbacks:

- Mere page count based methods are quite crude
- A text snippet may be quite long
- Word taxonomy may depend upon various heterogeneous aspects

- Processes like concept identification and relation identification are perception dependent

All these drawbacks have been the motivating factors for the proposed work as described in Chapter III.

2.4 LITERATURE SURVEY FOR ENTITY RESOLUTION

An *Entity* refers to a place, person, thing, event or abstraction having a distinct and separate existence from other instances of similar attributes e.g. *The Times of India, The Hindustan Times, Kabhi-Kabhi, Dilwale Dulhaniya Le Jayenge, i20, Santro Xing* etc. An entity may be referred with a list of formal and informal alternative names e.g. *TOI* and *Times of India* refer to the same entity. Similarly, *Tere Bin Laden-2* and *Tere Bin Laden: Dead or Alive* are not different. The references to an entity may be local or global depending upon the context of the underlying domain. These references are normally informal and cannot be resolved using the lexical resources.

Moreover, the alternative references are dependent upon the heterogeneity of the web spread over the various factors such as geographical domain, linguistic domain, education domain, slang terms etc. It is quite possible that a particular slang term or other term may serve as entity synonym for two or more entities e.g. the term *Godfather* when searched on the web may normally refer to the movie *The Godfather* but it is quite possible that it may be referring to *Godfather's Pizza* or *The Godfather Collection.*

Entity synonyms are important ingredients of current web search as major part of the web search is related to movies, events, monuments, books, players, actors etc. To find the appropriate entity through a particular entity synonym is known as *entity resolution.* Only way one can create a list of synonyms for an entity, usable in web search, is through the web content analysis.

Initial efforts to gather entity synonyms were based upon semantic knowledge and name aliases for most the prominent entities referred in knowledge bases like Freebase [98] and Wikipedia [99]. To collect valid entity synonyms from Wikipedia, redirection pages and disambiguation pages are used.

Following is the literature study in the area of Entity Resolution:

(a) Strube and Ponzetto [100] have talked about retrieving of entity synonyms using Wikipedia. Their work considers two strings to be entity synonyms if their Wikipedia category is same. The problem with this approach is that the size of Wikipedia is much smaller than the web and is limited to prominent entities only. Thus, the approach fails to take up the general purpose common entities.

To find the global and diverse synonyms of an entity, the vast and diversified extent of web can be the ideal source. Therefore, efforts are made to find web based empirical methods to generate the entity synonyms. Let us take a look on some of these efforts:

(b) Hu et. al. [101] used the redirection relationship between titles of the articles to find out entity synonyms. The approach suffers from the limitation of *title only concept* without taking into the account the page content as a whole.

(c) Chaudhury et. al. [102] have used the web search to find out the entity synonyms of a given entity name. Their work is limited to only those synonyms which are substrings of the entity name under consideration.

(d) Malekian et. al. [103] in their work have tried to convert a query into other forms using some features like word reordering, application/addition of modifiers, capitalization of alphabets etc. The work is not directly dealing with entity synonyms but is a contribution to the field of entity resolution in the sense that partial / inexact/ incomplete query can be handled through the system. For example, the queries like *toi, time of, the times of, time of india, TOI* can be converted to entity *The Times of India*.

(e) Some researchers [104-107] have tried to find out the entity synonyms using the reconciliation process in the databases. They have used divergent references of the same real world entity in separate or similar databases as entity synonyms.

The major problem with above citations is that they are unable to take into account the massive and heterogeneous content of the web. Moreover, in most of the cases, the availability of candidate reference is a priori requirement which should not be desirable. Actually the domain of entity resolution requires the automated generation of synonym candidate references from the web covering its vast and heterogeneous profile. These candidate references can then be pruned to create a set of credible entity synonyms.

(f) One such work has been published by Tao Cheng et.al [1]. In [1], Tao Cheng et.al. proposed a method based upon search data and click data to find the set of entity synonym. They have defined two sets A and L. The set A contains a set of tuples <q, p, r> wherein r denotes the relevance score of a web page p for the query q. The set L contains a set of tuples <q, p, n> wherein n denotes the number of times a user click on page p after issuing query q. The set A finds out the relevance relationship between the query and web page as observed by a search engine. The set L finds out the relevance relationship between query and webpage as observed by search engine users. Based upon this data, they have defined two ratios namely *Intersecting Page Count (IPC)* and *Intersecting Click Ratio (ICR)* which measure the strength of relationship between the two candidate strings based upon their surrogates pages identified and actually used (clicked). Larger the values of these ratios for a pair of query strings: more likely they are entity synonyms. The major achievement of the work is their pioneer effort to find the entity synonyms in automated manner using web query and search data. The major limitations of the work are:

- Click log sparsity problem that occurs when a query is asked by very few users and the clicked documents are even lesser.
- Inability to make a distinction between entities related to different concepts and classes e.g. *Oracle 10i* and *Oracle 10i tutorial* may be assumed as entity synonyms though they are referring to different concepts.
- Results are static, domain dependent and cardinality of the synonym set was quite less.

(g) Kaushik Chakrabarti et.al. [108] proposed a method to overcome the problems of click similarity [1] and document similarity [109]. Their work is based upon

the construction of a Pseudo-Document based upon the collection of all the tokens from all the queries that clicked on a document *d*. For this purpose, a query log is maintained for a time period and concept of reflexivity (synonym of self) and symmetry (*a* synonym *b* means *b* synonym *a*) is used. To remove the ambiguity, they have used the criteria of concept class (synonyms should refer to same concept) and auxiliary evidence (clicked documents). A pseudo document similarity function ensures the higher recall without dropping the precision.

(h) Srikantiah et. al. [110] proposed a mechanism to find the synonyms from the web on the basis of inbound anchor text. They have used Search Engine Result Pages (SERPs) to find candidate synonyms of individual keywords. The technique is scalable and can be applied to dynamic, domain independent data of unstructured web. The synonyms in their case are not entity synonyms but can be adapted to find out the entity synonyms. Their work has been a motivation for the work proposed in this thesis.

(i) Xiang Ren et.al. [111] proposed a new method of finding the entity synonyms that adopt a "structured" view of an entity by considering not only its string name, but also other important structured attributes. The approach is different from other contemporary methods based upon query log. The approach takes into account the structured view of an entity instead of abstract view related to string name. The work uses a graph based data model involving synonym candidates, web pages and keywords and their interaction relationship in the graph. The drawback of the work is its offline nature and a priori requirement of candidate synonyms.

2.4.1 Limitations of Existing Entity Synonyms finding Methods

The following limitations were identified while discovering entity synonyms using existing approaches:

1. No Lexical support as in case of word synonyms.
2. Entity references may vary with the global and local reference.
3. Synonym set generated through existing methods are not rich and global. They are unable to take into account the massive and heterogeneous content of the web.

4. Candidate synonyms are not generated by considering the contexts.
5. In many cases, the output is limited to only those synonyms which are substrings of the entity name under consideration.
6. In many cases, availability of candidate reference is a priori requirement which is not desirable.
7. There is no method for defining an index to assess the quality of synonyms generated.
8. Most of the approaches fail to take up the synonyms for general purpose common entities.
9. Some of the existing methods work for structured data and it cover the structured web queries with good precision. However, they are not applicable to dynamic and unstructured data i.e. WWW.

These challenges motivated us to propose a novel scheme for generating entity synonyms which is capable of working in a dynamic, online environment and it is not domain specific. The detail of the proposal is available in Chapter IV of this thesis.

2.5 LITERATURE SURVEY FOR RESOLVING CONCEPTS IN THE QUERY

A concept is an abstraction for which the intended set of entities needs to be identified. It is abroad idea generalized from its set of instances e.g. *Bird, Actor, Books, Movie* etc. The biggest challenge lies when a query contains concept(s) which has to be resolved through its meaningful set of instances as it requires awareness about worldly knowledge and the associations within. Consider the query *Best Universities in Europe.*

The easier part of the query is to list all the universities lying in the cities of Europe which only requires worldly knowledge but the difficult part is to decide about *Best Universities* in the absence of any predefined criteria in the query text. Similarly, we can consider queries like *Large Software Companies in Asia, Famous Bollywood Actor and Famous Musician etc.*

The handling of the concept is the most challenging task as its resolution requires the understanding and usage of worldly knowledge keeping in view the underlying

associations. The worldly knowledge is too vast to be comprehended and moreover becomes ambiguous, inconsistent and uncertain at many places.

The process of identification of entities associated with a concept in a particular context is known as *concept resolution* or *instantiation.* Concept resolution means to make proper instantiations for the concepts under considerations. It requires:

- Providing the machines/systems the access to large knowledge base related to common sense vocabulary
- Enabling the machines to use this knowledge in an unambiguous manner

Both of these are challenging tasks and can't be executed to perfection. But efforts can be made to accomplish this in a quite appropriate manner.

Enabling the machines, to have the common sense knowledge relating to this world, had long been the goal of the computer scientists. Initially, the interest was academic, relating to the development of intelligent systems covered under the domain of Artificial Intelligence. For this, efforts have been made by many researchers by the creation of manual taxonomies and ontologies. These efforts include the FreeBase (Bollacker et.al., 2008)[98], WordNet (Fellbaum, 1998)[97], Wiki Taxonomy (Ponzetto et.al.,2007)[112], Cyc (Lenat et.al., 1989)[113] , YAGO (Suchanek et.al., 2007)[114], KnowItAll (Etzioni et.al., 2004)[115], TextRunner (Banko et.al., 2007)[116], OMCS (Singh et.al., 2002)[117], NELL (Carlson et.al., 2010)[118]and DBPedia (Auer, 2007)[119] etc. Most of these taxonomies have been manually curated and contain limited number of concepts.

The number of concepts in the WikiTaxonomy, YAGO and Cyc are between 0.1-0.5 million, while in FreeBase, WordNet, DBPedia and NELL, their number is in thousands. When one takes into account the volume of common sense knowledge associated with this world, these numbers seem to be extremely small. In the practical applications requiring worldly knowledge (like web search), these resources prove to be quite inadequate. Therefore, a need has long been felt to develop the huge taxonomies and ontologies based upon the web pages in order to:

- Cover large number of concepts and their instances.

- Cover the heterogeneity and versatility of the web.
- Deal with the probabilistic or partial relationship between the concepts and entities in consideration.

One such effort has been in the form of *PROBASE* in Wu et.al. (2012)[120, 121], Song et.al (2011)[122], Lee et.al. (2011)[123], Lu et.al (2012)[124], Liu et.al. (2012)[125] and Wang et.al. (May 2011)[120, 121], a project carried out by Microsoft research Asia which includes more than 2 million concepts and their associated entities. The biggest strength of the PROBASE lies in its two characteristics.

1. The taxonomy has been derived from the web, therefore it involves the actually used concepts by the people worldwide involving all sorts of heterogeneity and slang terms.
2. The size of the taxonomy is huge and contains very large number of general terms which is much bigger (by one order of magnitude) than its nearest competitors.

The PROBASE by Wu et.al.(2012)[120] includes a large number of concepts and a very large number of associated instances e.g. the concept *actor* has been associated with more than 3000 instances. There are the many concepts that have been associated with hundred thousand instances making it quite difficult to associate the proper set of instances to the corresponding concept. Efforts in this direction include the works of Wang et.al. (2012)[125], Egozi et.al. (April 2012)[126] and Sendhil et.al. (2010)[127]. These works lack depth and operate at quite a surface level. These are discussed below:

(a) The work proposed by Wang et.al. [26] considers short text as "Bag of Concepts" without taking into consideration the document as a whole.
(b) Explicit Semantic Analysis proposed by Egozi et.al.[126] uses relatedness analysis based on Wikipedia but neglects the context of words and cannot exactly determine the desired sense of an ambiguous word.
(c) The work proposed by Sendhil et.al.[127] deals with construction of personalized page view graph for small scale search which is limited to an individual only.

(d) Fonseca et.al. (2005)[128] generated and organized concept hierarchy from the stored document sets and used it for query expansion purposes with a view to improve precision.

(e) Lu et.al. (2017)[129] used TREC-VID 2015 (multimedia event detection system) for handling complex concepts in the user query. Their system detected large number of concepts using pre-trained concept detectors for textual-to-visual relation. The problem with this system is the restriction of using only multimedia event detection system for handling the concepts.

(f) Metzler et.al.(2007)[130] proposed a new mechanism known as latent concept expansion for expanding the term concepts for tasks such as query suggestion and query reformulation.

(g) Boucennaet.al. (2016)[131] proposed concept-based semantic search for outsourcing the data over cloud after encrypting it. The major restriction using this approach is, all data must be encrypted before being outsource into the cloud and additional overhead occurs for the purpose of encryption & decryption.

This study of literature helped us in identifying following facts about concept instantiation:

- Manually curated worldly knowledge sources such as NELL, Wikipedia, Freebase, DBPedia are insufficient to fill the requirement of the web search engines.
- There is a need to have a worldly knowledge source with the ability to handle all sort of heterogeneous and multidimensional knowledge pertaining to this world.
- PROBASE is an effort in this direction and is publically accessible on the https://concept.research.microsoft.com/Home/Introduction
- Google Humming Bird principle indicates that the user's geographical location, Browsing History and other such parameters can be used to cater the interest of individual user.
- In the WebPages, a lot of concepts are described in the form of slang terms, which are not defined in the lexical sources e.g. the concepts biggie, bigwig, big-wig, heavyweight etc.

- Google has shifted to knowledge graph based search from keyword based search.
- A lot of time is wasted by the search engine if the same query can be interpreted in multiple manners e.g. the string “New York Times Square Problem” can be interpreted as

  New York Times and square problem
  New York and Times Square problem

After identifying the above facts about the present state of affairs regarding the concept instantiation process, following needs were identified to incorporate precision in concept-instantiation process.

- Availability of a credible worldly knowledge source with huge amount of knowledge relating to heterogeneous aspects of the real world.
- Availability of an index in the knowledge source to distinguish between the strength of various candidate instances of a particular concept.
- Analysis of various available contexts in the form of geographical location of the user, past browsing history, IP address and other such trivial information can be of immense help in going for appropriate instantiation as per the need of the user.
- A meaningful parsing of the query text can reduce a lot of burden on the web search engine. So, there should be some appropriate text writing mechanism for resolving between multiple possible meanings of the query text.
- Slang terms should be clubbed with their closest possible formal terms.

Keeping all these aspects in view, a mechanism has been proposed for resolving a concept to its appropriate set of closest instances, using PROBASE, in the presence of available contexts such as IP address, browsing history etc. The details of the mechanism have been discussed in Chapter V of this thesis.

2.6 FUZZY LOGIC

In the preceding text, we have been talking about the need for an index to measure the extent of similarity between various candidate synonyms of a given word/ entity. The

measuring of the index is normally not very precise and is dependent upon various factors used in the synonym identification process. Therefore, two synonyms with not much difference in their similarity indices can be safely considered in the same linguistic category as used in *Fuzzy Sets*. This can also help in designing the automated applications related to web search using *Fuzzy Rule Base* (FRB) system. In this section, we take a brief look at the various aspects of *Fuzzy Sets and Logic.* The architecture of a FRB system for making the required inference has also been discussed.

Fuzzy logic [132] is a form of multi-valued logic derived from fuzzy set theory to deal with approximate reasoning. In contrast to "crisp logic" that has binary values (1/0 or true/false), a fuzzy logic variables may have the value between 0 and 1indicating the degree of truth, 0 being completely false and 1 being absolutely true and other being partial truth and partial false. Similarly, a Fuzzy set is a superset of conventional (Boolean) set wherein an element can partially belong to a set.

2.6.1 Sets and Logic

A set is a well-defined collection of objects wherein an object either belongs to the set or it does not. This concept is mathematically defined by using a characteristic function, *f(A ,x)*

$$f(A,x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$$

where *A* is a set and *x* is any object. The problem with this type of classical or normal set is that it only talks about complete belongingness or un-belongingness and does not deal with partial belongingness.

To cover this issue, concept of fuzzy sets was introduced by Lotfi A Zadeh [209]. In a fuzzy set, there is a membership function μ that indicates the extent of belongingness i.e. $\mu_A(x) = [0,1]$

where 0 indicates no belongingness, 1 indicates total belongingness and ($0<\mu<1$) shows partial belongingness. The membership can also be expressed as A(x).

2.6.2 Defining a Fuzzy Set

Fuzzy sets represent linguistic concepts such as *very small, small, medium, large, huge etc.* with their interpretation in a particular context expressed through linguistic variables. The states of linguistic variable can be defined in terms of base variable having real number values within a specific range. A base variable is a variable in the classical sense, exemplified by any physical variable (e.g. temperature, pressure, speed etc.) as well as any other numerical variable (e.g. age, interest rate, salary etc.). Each linguistic variable is fully characterized by a tuple (v, T, X, g, m) where

v→ Base variable
T→ set of linguistic terms of v that refer to a base variable whose values range over a universe set X.
g→ syntactic rule for generating linguistic terms
m→ semantic rule that assigns to each linguistic term t∈T

To illustrate the definition of fuzzy set, let us take the example of human age.

Base Variable: Human Age
Domain of discourse: (0,125) yrs.
Linguistic terms for Fuzzy sets: Child, Young, Middle Age, Old, Very Old.
Semantic Aspect: Linguistic expressions or names of Fuzzy sets should be meaningful.
Syntactic rules: Fig. 2.1 shows the syntactic interpretation of various fuzzy sets.

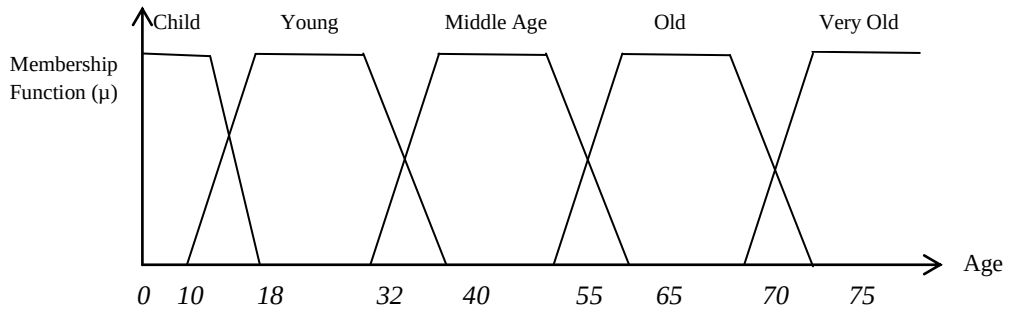


Fig. 2.1 Representation of a Fuzzy Set for base variable age

In this case the set S (the universe of discourse) is the set of people of different age group. In the above diagram five different fuzzy sets are defined such as *Child, Young, Middle Age, Old* and *Very Old* etc. In order to define a fuzzy subset YOUNG, which answers the question like "to what degree is person x young?", we have to assign a degree of membership in the fuzzy subset YOUNG to each person in the universe of discourse. The easiest way to do this is with a membership function based on the person's age. To define the membership function for the fuzzy set young, four possible cases can be there. The value of membership grade is zero when the age is less than 10 and greater than 40, its value increases from 0 when age is between 10 to 18, its value decreases from 1 when age lies between 32 to 40 years, the value is exactly one when age lies between 18 to 32.

2.6.3 Terms Associated with Fuzzy Set

1. **Membership:** It indicates the extent of belongingness of an element to a particular fuzzy set. Its value lies between in interval [0,1]. Larger value of μ indicates more belongingness to the set.
2. **Support of a Fuzzy Set:** Support is that numeric range wherein the μ>0 e.g. sup (child)=(0,18).
3. **Height of a Fuzzy Set:** Height of a fuzzy set is the highest value which μ attains within its support.
4. **Normal Fuzzy Set:** When the height of fuzzy set is 1, it is known as normal fuzzy set.
5. **Subnormal Fuzzy Set:** If design of fuzzy set is such that its membership value never reaches '1', then it is a subnormal fuzzy set. A subnormal fuzzy set has height less than one.
6. **Continuous Fuzzy Set:** In an continuous fuzzy set, the values of the element can be real numbers i.e. the domain of fuzzy set is real **e.**g. Sup(child)=(0,18) represents a continuous fuzzy set whose element can have any real numbers having values from 0 to 18.
7. **Discrete Fuzzy Set :** A discrete fuzzy set has discrete domain wherein the element

   under the consideration can have any discrete and quantized value.

   e.g. $A(x)=\frac{0.2}{4}+\frac{0.3}{5}+\frac{0.7}{6}+\frac{1}{7}+\frac{0.6}{8}+\frac{0.4}{9}$

8. **Alpha Cut of a Fuzzy Set**: An α-cut indicates a numeric range in the support of fuzzy set A, wherein $\mu \geq \alpha$ it is indicated by A^{α}.
9. **Strong alpha cut**: Support of a fuzzy set can be defined in the form of strong α-cut. A strong α-cut indicates the numeric range within the support of the set wherein the $\mu > \alpha$ it is indicated by $A^{\alpha+}$.
10. **Fuzzy Rule Base (FRB):** A FRB is a set of deductive reasoning rules which deal with the inference relationship which can be used for decision making process in a control system/ expert system. These rules are stored in the knowledge base of a control system and are selected for the usage in a particular context depending upon the user's requirements. Some example fuzzy rules are as follows:

    If *temperature is very high* then *valve opening is large*.
    If *pressure is medium* then *valve opening is very low*.

The Fig. 2.2 shows a Fuzzy Control System that uses FRB to control a process. The working of the fuzzy control system is available in various text books. This system was adapted to automate the web search process as shown in the Fig. 2.3.

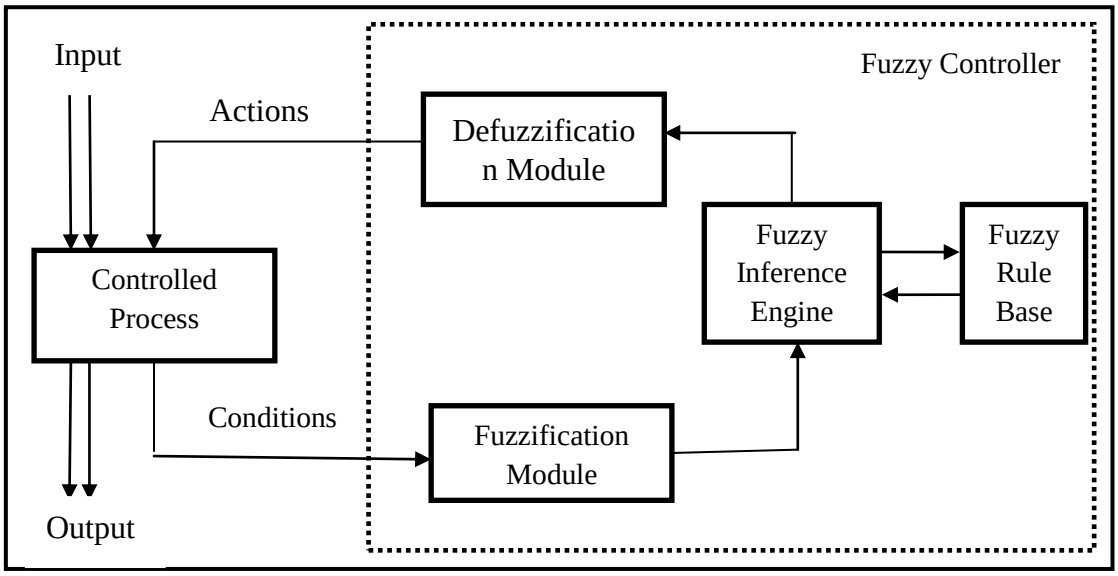


Fig. 2.2 Fuzzy Control System

The Fuzzy Control System (FCS) shown in the Fig. 2.2 was adapted for search engines in order to make the decision as shown in Fig. 2.3 for purpose of query expansion and rephrasing. The adapted system is as shown in the Fig. 2.3

Some example rules used in the query expansion/ rephrasing process are as follows (Here W1 is the original word and W2 is its synonym and similarity index represents the similarity between them in the form of fuzzy set):

If *similarity index is perfectly_similar* use W2 in place of W1 for various purposes like query expansion, query reformulation, word sense disambiguation etc.
If *similarity index is quite_similar* use W2 in place of W1 for query expansion, query reformulation, word sense disambiguation etc.

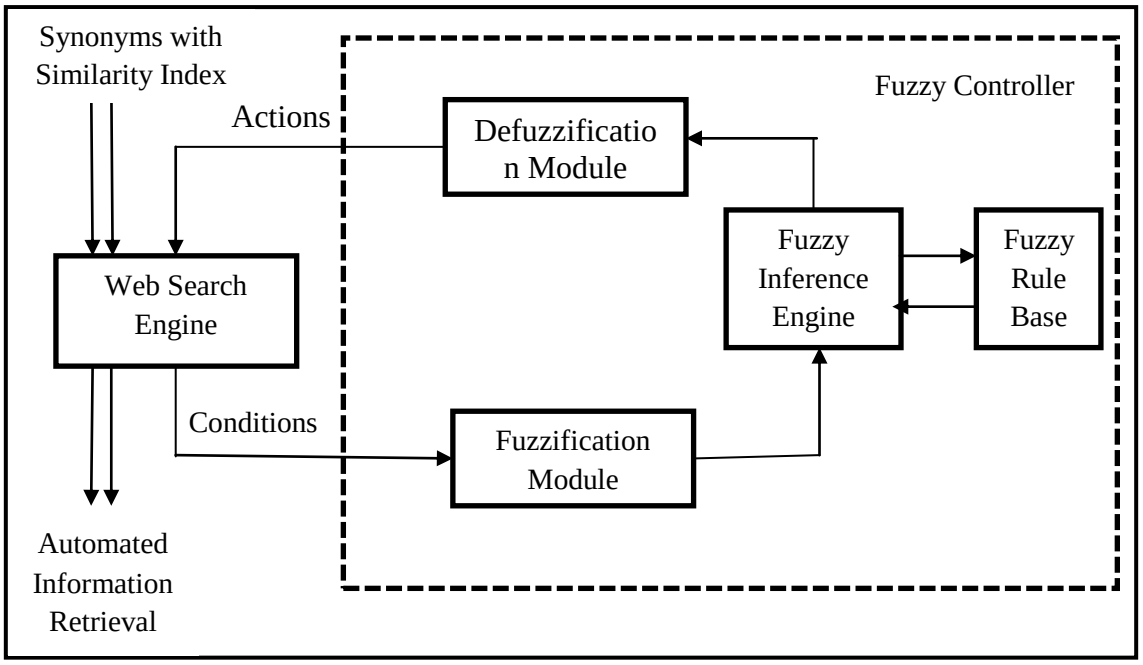


Fig. 2.3 Modified Architecture of Fuzzy Control System

After taking up the details of the study carried out in this chapter, we take up the details of the proposed work in the subsequent chapters.

CHAPTER III
SYNONYM RESOLUTION FOR ATTRIBUTE COMPONENT IN QUERY

3.1 INTRODUCTION

In today's digital world, for any topic of interest, information can be retrieved from World Wide Web (WWW). But due to enormous size of the Web, it is not an easy task to retrieve exact information. To overcome the problem, variety of search engines are available online which provide an easy to use interface wherein a user can express his/her information requirements in the form of query. A search engine has to ensure the retrieval of desired information content as per the requirements of user. Query being a very *short piece of text* has to be critically evaluated for its unambiguous and exact meaning. The various pages available on the WWW which can be relevant to the user's query may not have exactly same matching words as used in the query. Thus, it is a requirement that the synonyms of the words used in the query be chosen in an appropriate manner so that more and more relevant pages included with non-relevant being discarded. This mandates that an appropriate mechanism be devised to create the synonyms of the words based on the huge real world data which is quite feasible in today's technological scenario. Also, it is desirable that the extent of similarity between a word and its synonym be quantified to enable the search engines in taking a rational decision for the purpose of query substitution and expansion. This chapter takes up this task and proposes a mechanism for finding appropriate synonym set for a given word and to express the extent of similarity between them using both numerical measures and fuzzy sets. The work has been carried out for *attribute* component of the query.

After the inception of the Web, need for the search of meaningful and appropriate synonyms has been felt like never before. The requirement is no longer mere academic but is necessity for appropriate and relevant information retrieval from the web. The manually curated lexical resources like WordNet [97], YaGo [114], Cyc [113] are unable to meet this requirement due to gigantic size and heterogeneity of web. Therefore, a lot of efforts have been carried out in the past to find out the word similarity for finding meaningful synonyms using web pages. These works have been

discussed in literature survey of this thesis in Chapter II. Before taking up the proposed work, we take a look on the limitations of these contributions.

3.2 LIMITATIONS OF EXISTING WORD SIMILARITY METHODS

The existing word similarity approaches suffer from the following limitations:

- Mere page count based approaches used in [10, 11] are quite crude.
- Text snippets is based approaches [133] are dependent upon choice of snippet length. Both short and long snippets have their own merits and demerits.
- Word taxonomy based approaches [95] depend upon various heterogeneous aspects.
- Processes like concept identification and relation identification are perception dependent [17].
- Word embedding models used in [134] are not much consistent and their applicability depends upon case to case.

These challenges motivated to direct the research in calculating the word similarity using real world data.

3.3 PROPOSED WORK

The semantic aspect of a query is essential while determining the information required by the user using available information retrieval tools. For retrieving documents, which are semantically similar to the query submitted by user, it is required that users' query be expanded to cover more terms which are semantically similar. A pure keyword based information retrieval system is unable to use the concept of semantic similarity thereby missing the major chunk of useful pages. The search engines, if do not take into account the context of a query and unseeingly use the online lexical resources, may lead to fetching of large number of undesired pages which are otherwise not essential. To defeat this weakness, it is required that query be expanded through meaningful semantic expressions using context set. This appropriate choice requires the precise knowledge about semantic similarity between original word in the query and the word to be substituted. But the precise measurement of the semantic similarity is a challenging task as most of words in a natural language have context

dependent meaning. Keeping this need in view, this chapter focuses on a novel approach which tunes the lexical resources wherein the derived synonyms of a word are also provided with their applicable context.

Since it is not possible to explore whole of the Web, therefore for finding the synonyms, help of multiple corpora, with wide range of genres and with each one containing huge set of documents were taken. The snapshot of different Corpuses used to extract context related to a word is shown in Appendix A.

The work presented here takes into consideration the attributes present in the query. The attribute in a query is a word that describes a particular quality of the subsequent word in the query. The subsequent word whose qualities are described using the attribute can be considered as its context. For example, if two or more attribute words have similar context set then they must be synonyms. The proposed work is based upon the following strategies:

- Identification of the context set of the attribute word under consideration using multiple corpora.
- Adaptation of existing similarity measures for computing the similarity between various context sets.
- Computation of similarity extent between the contexts sets using different adapted similarity measures.
- Normalization of similarity scores computed through various adapted similarity measures.
- Designing of fuzzy sets to give the computed similarity a meaningful linguistic expression.

The proposed work can contribute to the web search technology and the linguistic world in following manner:

- Fuzzy sets generated can be used in the creation of the Fuzzy Rule Base (FRB) for automated use of computed similarity index for applications like web search.
- The new generation lexical resources can also be augmented with the applicable similarity index while defining the synonyms of a word.

- The context set of a word can also be defined to ensure its proper usage.

Thus, the attribute synonyms promise to deliver more precise results for a semantic search rather than keyword based search.

Before going for actual process, let us take up the details of the online resources used in the proposed work.

3.3.1 The Resources

Various online lexical resources and the corpora have been utilized as knowledge base for the proposed work, the details of which are given as under.

(i) WordNet

WordNet is one of the most influential online lexical resource developed by Miller et.al.[97] at Princeton University. It has combined features of both dictionary and thesaurus, based upon psycholinguistic theories of human lexical memory. It consists of set of English nouns, verbs, adverbs and adjectives organised into synsets and having various lexical-semantic relations between them. In this work, this resource is used to retrieve the set of synonyms for substituting the input word used in the query.

(ii) Corpora

To identify the contexts related to a word, four different corpora are considered namely Coca Corpus [135], BNC Corpus [136], Wikipedia Corpus [137] and GloWbE Corpus [138]. The size of each corpus along with the other details is shown in Table 3.1. The reason for choosing the above corpora is their ability to allow search on the basis of word, phrase, part of speech and synonyms. For the purpose of context identification, the size of text window is taken as 2 because a shorter window ensures the proper relevance of context. The input word is searched into the corpus to extract context related to a word. These contexts are then used to find semantically similar words.

Table 3.1 Details of Corpora for Consideration

Corpus	Genres	No. of words	Type(s)of documents
Coca Corpus	spoken, fiction, popular magazines, newspapers and academic texts.	520 million	Text
BNC Corpus	spoken, fiction, magazines, newspapers and academic	100 million	Text
Wikipedia Corpus	documents related to microbiology, economics, basketball, Buddhism, or thousands of other topics.	1.9 billion	Text
GloWbE Corpus	Any type of data related to newspaper, magazines and academic.	1.9 billion	Text

The relational database design of these corpuses allows complex queries to be executed in two or three seconds. These corpuses are used to extract context related to a word and helps to build next generation lexical resource. The upcoming section throws light on the foundation of the proposed work.

3.3.2 Proposed Synonym Resolution Approach for an Attribute using Contexts

The main objective of the work is to propose a synonym resolution method for attributes in a query based upon the immediate context of the said attribute in various corpora. The outcome of the work includes context set identification for a word and computation of an index indicating the extent of semantic similarity between a pair of words. The computed index has been fuzzified into a fuzzy rule base for the purpose of automation and its usage into the web search engines and other such applications. The proposed architecture for finding the semantically similar word of the attribute word in consideration is shown in Fig. 3.1.

The proposed approach uses WordNet and exploits various corpuses to identify the context set for the words under consideration. The reason for these choices is the extensive coverage of almost every branch of knowledge by them and the volume of available data. Initially the input attribute word is searched in the WordNet for extracting the set of available synonyms. Thereafter, four different corpuses have been used to identify the context set for both the word under consideration and its chosen synonym. The list of all possible contexts are taken out by considering the

union of contexts taken from all the four corpora. The list of most commonly used contexts is taken out by considering the intersection of these contexts. Now, the extracted list of synonyms is checked for similarity with the input adjective word. In the literature, WebJaccard, WebOverlap and WebDice are different types of standard indices used to compute similarity index between input word and its synonym word on the basis of page-count. All these indices have been modified in the proposed algorithm to use the context set cardinality instead of page count. Now, one can choose an appropriate index and use it for fuzzification. Our preferred choice is Modified WebJaccard as it takes into account both union and intersection aspect while calculating similarity.

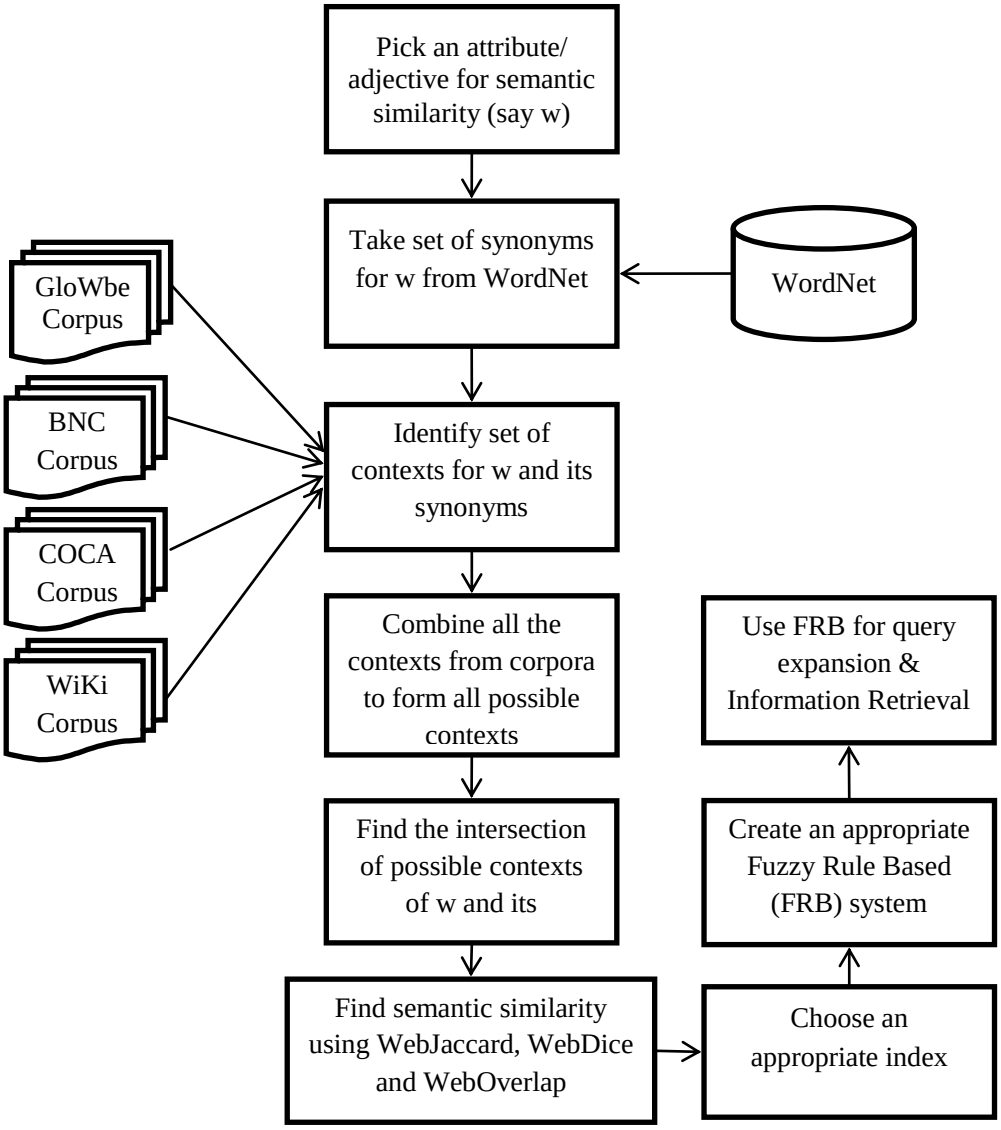


Fig. 3.1 Proposed Architecture for finding the Semantically Similar Word

To enable the automated tools in making the intelligent decision on the basis of the additional information created in the work, the similarity index has been fuzzified through the design of appropriate fuzzy sets. This will help the information retrieval system in following manner:

(a) Minor differences arising out of computation due to the choice of search engine, set of documents, dialects etc. are eliminated, as two close values would normally be residing in the same fuzzy set.
(b) The fuzzy sets created in such a manner can be used for intelligent decision making by creating a Fuzzy Rule Base (FRB) that can be run on an appropriate fuzzy inference engine.

3.3.3 Creating Semantic Similarity Index and Applying Fuzzy Rule Base (FRB)

Beginning with the details of proposed work, the following algorithm has been designed:

(a) **Algorithm for finding the semantically similar word**

Input: The word *w* (adjective), *n* sets of word corpus

Output: Set of semantically similar words of *w* along with similarity index, Fuzzy Rule Base (FRB)

Step1: Choose a candidate word for finding semantic similarity, say *w*

Step2: Choose a set of *n* corpora.

Step 3: For each corpus *i*, identify set of contexts for *w*, C_i

Step 4: Compute $C_{pw} = C_1 \cup C_2 \cup \ldots \cup C_n$ // *Possible list of contexts for w*

Step 5: Compute $C_{cw}=C_1 \cap C_2 \cap \ldots \cap C_n$ //*List of most commonly used contexts for w*

Step 6: Augment C_{pw} and C_{cw} in lexicon along with the word *w*.

Step 7: Let *S* be the set of synonyms for *w* obtained from a lexical resource, say WordNet.

Step 8: for each $x \in S$, Compute $C_{px}$and C_{cx} using step 3-5 //*List of possible and commonly used contexts for x.*

Step 9: Compute similarity index between *w* and *x* using WebJaccard co-occurrence measure. // Context set cardinality is used instead of page count.

Modified_WebJaccard (*w*, *x*)= $|C_{pw} \cap C_{px}| / |(C_{pw} + C_{px} - C_{pw} \cap C_{px})|$

Step10: Normalize the computed co-occurrence to a range $[r_1, r_2]$.

Step11: Use $[r_1, r_2]$ as domain of discourse to create fuzzy sets expressing the extent of similarity between the words *w* and *x*. Let the fuzzy sets be: *not_similar*, *poorly_similar*, *medium_similar*, *highly_similar*, *extremely_similar*.

Step12: Create an appropriate Fuzzy Rule Base (FRB) to implement the computed similarity in automated manner.

Step 13: Use FRB for query expansion and information retrieval.

Step 14: Stop

(b) **Explanation and implementation of proposed algorithm**

The proposed algorithm was implemented using a very commonly used word *beautiful*. The explanation behind picking the word *beautiful* is because of its commonness and availability of its large of synonyms applicable to different varieties of contexts. Its available synonyms were discovered using the available lexical resource *WordNet* thereby creating the set *S*. With the goal of context identification, the size of text window (i.e. the maximum distance between the focus word and its contextual neighbours) was taken as 2 (two) due to the fact that a shorter window ensures the proper relevance of the context. In addition, since the word is of adjective type, a window of size 2 is quite appropriate. For different sort of words, the size of the window can be adjusted. The calculation of C_{pw} (possible list of contexts for word *w*) and C_{cw} (list of most commonly used context for the word *w*)using multiple corpora guarantees the elimination of bias that may occur in a specific corpus.

To apply the algorithm, few synonyms of *beautiful* were taken from the WordNet namely: *Pretty, Lovely, Gorgeous, Glorious, and Stunning.*

The extracted list of synonyms was checked for similarity with the input adjective word. The different corpora (here four) were used to identify the contexts related to a given word and its synonyms. Then list of all possible contexts were taken out by considering the union of contexts from all the four corpora. The list of most commonly used contexts was taken out by considering the intersection of contexts using the four corpora. In the literature, *WebJaccard, WebOverlap* and *WebDice* are different type of standard indices are used to compute similarity index between input word and its synonymous word on the basis of page-count. All these indices have

been modified in the proposed algorithm to use the context set cardinality instead of page count. It is for this reason that they have been named as Modified_WebJaccard, Modified_WebOverlap and Modified_WebDice respectively.

The computed index has been normalized to the range [0,100] through the use of a normalization factor. The purpose of normalization is to create a relative standing between the various candidate synonyms through a standard level of parity e.g. percentage. The so computed normalized index can also be used as the membership indicator in the corresponding fuzzy set. The range [0,100] is then used as a domain of discourse to create fuzzy sets expressing the extent of similarity between two words through linguistic expressions. The fuzzy set framework created in such a manner leads to the creation of a Fuzzy Rule Base (FRB) that can be used for automated query expansion and information retrieval through the usage of computed normalized similarity index.

The major issue here was the calculation of normalization factor. The empirical calculation of this factor may slightly vary with the perception of the designer. In our case, it is considered that two different words are almost similar (assuming that they cannot be exactly similar) at the level of 100 if their context set is at least 80% same as that of the smaller context set and sizes of their context sets do not differ by 25%. Of course, the designer of the automated system can choose different values as suited to their application and the precision requirement.

To illustrate the normalization process, let us take two words w1 and w2 with cardinality of their context sets as $|C_{w1}|$ and $|C_{w2}|$.

Let $|C_{w1}| = 100$, $|C_{w2}| = 125$
Here difference in cardinality of context set is 25%
It is assumed here that context set matching is 80% of the smaller context set i.e.
$|C_{w1} \cap C_{w2}| = 80$
Here, $|C_{w1} \cup C_{w2}| = |C_{w1}| + |C_{w2}| - |C_{w1} \cap C_{w2}| = 145$
Expected normalized similarity value=100
This data leads to values of different indices as follows:
Modified_ WebJaccard= 80/145=0.55
Modified_ WebOverlap as 80/100=0.80

Modified_ WebDice as 160/225=0.71.

The normalization factor under the circumstances for these indices will be 100/0.55=182, 100/0.80=125 and 100/0.71=141 respectively subject to maximum of 100. The so computed normalization factor has been used in the Tables 3.3 to 3.9 of this chapter to compute the normalized similarity index.

(c) Example Context Sets

Given below are possible and commonly used context sets for the example word 'beautiful' and its synonyms.

List of possible contexts for word Beautiful:

C_{pw}={ area, art, baby, bay, beach, black, blonde, blue, body, book, boy, bride, building, children, church, city, cloth, color, country, countryside, creature, dark, daughter, day, dreamer, dress, evening, eyes, face, family, female, fish, flower, game, garden, girl, hair, head, home, house, image, island, job, lady, lake, landscape, liar, life, little, maiden, man, mind, morning, mother, mountain, movement, music, name, natural, new, night, object, old, park, part, person, photo, picture, piece, place, post, princess, red, river, room, scenery, setting, sight, singing, skin, smile, song, soul, sound, south, spring, story, stranger, summer, surrounding, temple, thing, town, valley, view, village, voice, way, weather, wedding, white, wife, woman, word, work, world, young}

Here $|C_{pw}|$=107

List of Commonly used contexts for word beautiful:

C_{cw}={baby, beach, city, country, daughter, day, face, flower, garden, girl, lady, person, place, scenery, thing, view, voice, wife, woman, young}

Here $|C_{cw}|$=20

List of Commonly used contexts for word Pretty:

C_{CX1}={girl, woman ,face, picture}

Here $|C_{cx1}|=4$

List of possible contexts for word Pretty:

$\mathbf{C_{PX1}}$={ baby, blonde, boy, busy, child, cloth, committee, dress, eyes, face, garden, girl, guardian, horse, lady, light, maid, picture, pink, poison, sight, soldier, solid, song, sweet, thing, village, visitor, woman}

Here $|C_{px1}|=29$

List of Commonly used contexts for word Lovely:

$\mathbf{C_{CX2}}$={bone, children, city, couple, daughter, day, eyes, face, family, garden, girl, guy, lady, man, night, person, place, sight, song, story, thing, voice, wife, woman}

Here $|C_{cx2}|=24$

List of possible contexts for word Lovely:

$\mathbf{C_{PX2}}$={afternoon, angel, area, article, ballad, bank, blog, blue, body, bone, book, boy, bride, bunch, butcher, carrot, children, city, cloth, color, colour, country, countryside, county, couple, daughter, day, dinner, dog, dream, dress, egg, evening, eyes, face, family, feather, feeling, flower, food, friend, garden, gift, girl, green, guy, hair, head, holiday, horse, hotel, house, idea, idol, image, kid, lady, land, lane, light, little, lock, lunch, man, meal, moment, music, name, night, old, party, person, photo, picture, piece, place, princess, reader, red, rita, room, rose, seat, setting, shade, sight, smell, smile, song, sound, spot, story, stuff, summer, surprise, tea, thing, thought, time, town, valley, view, voice, war, way, weather, white, wife, woman, word, world, young}

Here $|C_{px2}|=112$

List of Commonly used contexts for word Magnificent:

$\mathbf{C_{CX3}}$={achievement, bird, building, city, collection, display, example, garden, house, job, old, performance, piece, scenery, sight, view, work}=17

Here $|C_{cx3}|=17$

List of possible contexts for word Magnificent:

$\mathbf{C_{PX3}}$={achievement, amberson, animal, architecture, art, away, backdrop, baroque, bastard, beach, beast, beauty, bird, black, blue, body, book, bridge, bronze, building, but, butcher, career, castle, cathedral, century, church, city, cliff, collection, contribution, coral, country, countryside, courage, court, creation, creature, day, desolation, dining, display, dog, dress, effort, entrance, event, example, experience, figure, five, four-poster, fraud, frigatebird, funeral, game, garden, gift, goal, golden, gothic, grey, hall, head, history, home, horse, hotel, house, job, journey, lake, landscape, library, life, man, mansion, marble, medieval, mile, montague, monument, moody, mountain, muraco, natural, new, obsession, old, opportunity, palace, panorama, park, performance, physique, piece, place, player, record, red, renaissance, response, ring, room, scale, scenery, season, second, set, setting, seven, show, sight, site, six, sound, specimen, spectacle, stained, state, stone, structure, sunset, surrounding, temple, thing, tomb, tree, victory, view, villa, vista, voice, way, white, woman, wooden, work, world, yankee, young}

Here $|C_{px3}|=141$

List of Commonly used contexts for word Good_Looking:

$\mathbf{C_{CX4}}$={ boy, couple, dog, face, fellow, female, girl, guy, man, person, woman}

Here $|C_{cx4}|=11$

List of possible contexts for word Good_Looking:

$\mathbf{C_{PX4}}$={ actor, african-american, american, animal, babe, baby, bastard, bird, black, blog, bloke, blonde, boat, body, boss, boy, breed, british, broad, brother, brunette, cabin, car, castored, chap, character, chestnut, chick, child, children, cloth, club, college, color, cook, cookware, corpse, couple, crowd, curve, daddy, danish, date, design, desk, detective, diagram, dish, doctor, document, dog, dreamer, drew, duck, dude, duke, european, face, family, feature, fellow, female, field, fighter, film, folk, football, foreigner, frame, friend, game, gentleman, girl, glass, graphics, group, guard,

gun, guy, hair, handset, head, horse, husband, image, kid, lady, living, male, man, manuscript, mary, movie, musician, officer, one, pair, patient, person, phone, photo, pig, playboy, product, professional, series, shot, stranger, sunshine, surgery, teacher, teenager, victory, view, voice, waiter, weather, white, woman, year, young, youth}

Here $|C_{px4}|=122$

List of Commonly used contexts for word Glorious:

$\mathbf{C_{CX5}}$={ battle, career, celebration, city, day, death, era, food, future, history, land, life, light, moment, night, revolution, summer, sun, thing, tradition, victory, voice, year}

Here $|C_{cx5}|=23$

List of possible contexts for word Glorious:

$\mathbf{C_{PX5}}$={ abandon, achievement, adventure, afternoon, age, amateur, appearing, army, ascension, battle, beach, bloom, blue, body, book, burden, career, cause, celebration, century, chance, chapter, church, city, climax, color, colour, company, country, countryside, day, dead, death, deed, display, empire, end, era, experience, eyes, fire, first, food, freedom, future, garden, goal, god, gospel, green, hair, heritage, high, history, human, king, land, leader, life, light, lord, love, memory, mess, minute, mission, moment, month, morning, mother, mud, music, mystery, name, nation, night, noise, opportunity, order, past, path, peace, period, place, player, power, prospect, red, return, revolution, ring, role, run, scenery, season, sense, sight, song, sound, spring, success, summer, sun, sunrise, sunset, sunshine, thing, time, tradition, victory, view, virgin, voice, war, way, weather, week, work, world, year, youth}

Here $|C_{px5}|=121$

List of Commonly used contexts for word Stunning:

$\mathbf{C_{CX6}}$={ array, beauty, collection, landscape, performance, photograph, piece, scenery}=8

Here $|C_{cx6}|=8$

List of possible contexts for word Stunning:

C_{PX6}={ about, accomplishment, ace, achievement, admission, album, amount, announcement, arcade, architecture, array, art, artwork, attack, beach, beauty, black, blonde, blue, book, cast, central, claim, clarity, coast, collapse, collection, color, conclusion, contrast, costume, countryside, creation, dark, day, debut, decision, defeat, design, detail, development, discovery, display, dot, double, dress, effect, election, end, evening, example, exhibition, fact, fall, fashion, film, fish, form, garden, girl, goal, good, graphics, guitar, hat, home, illusion, image, impact, lake, landscape, lap, light, line, location, look, loss, military, moment, mountain, move, natural, news, number, opening, panoramic, performance, photo, photograph, picture, piece, place, portrait, presentation, record, result, reversal, scenery, sea, season, series, set, sight, solo, souvenir, speed, start, statement, story, strike, success, surrounding, thing, turn, variety, victory, video, view, visual, voice, volley, white, woman, work, young}

Here $|C_{px6}|$=125

All the mentioned sets are utilized during the implementation and analysis of proposed approach.

3.4 RESULT AND ANALYSIS

Table 3.2 represents the computation of various indices based upon the extracted information. Results have been normalized using the normalization factor described in previous section of this chapter. The computed results have been compared with UMBC toolkit, [139] in Tables 3.3 to 3.9. The toolkit uses statistical method that is based on Latent Semantic Analysis (LSA) and distributional similarity. The whole process is automated and can be trained using different corpora. The technique used in this toolkit assumes that the semantics of a phrase is compositional on its component words. The concept and relation similarity can be found for noun, verb, adjective and adverb using either refined Stanford Webbase corpus [140] or LDC English Gigaword corpus [141].

Table 3.2 Computation of different Indices

	Corpus	**C_{pw}**	**C_{px}**	**C_{pw}**	**C_{pw}**	**Modified**	**Modified**	**Modified**	**UMBC**
Beautiful									
Pretty	*COCA*		29	16	120	0.133	0.552	0.235	40
Lovely	+BNC+*WIKI*+		112	60	159	0.377	0.560	0.548	80
Magnificent	GloWbE		141	47	202	0.232	0.439	0.377	60
Glorious			121	34	194	0.175	0.318	0.298	30
Good_Looki		107	122	28	201	0.139	0.262	0.245	Not
Stunning			125	35	197	0.177	0.327	0.302	50
Pretty									
Beautiful	*COCA*		107	16	120	0.133	0.552	0.235	40
Lovely	+BNC+*WIKI*+		112	15	126	0.119	0.517	0.213	50
Magnificent	GloWbE		141	6	165	0.036	0.207	0.070	Not
Good_Looki			122	10	141	0.071	0.345	0.132	30
Glorious			121	6	144	0.042	0.207	0.080	Not
Stunning		29	125	9	145	0.062	0.310	0.117	Not
Lovely									
Beautiful	*COCA*		107	60	159	0.377	0.560	0.548	80
Pretty	+BNC+*WIKI*+		29	15	126	0.119	0.517	0.213	50
Magnificent	GloWbE		141	33	221	0.149	0.295	0.260	50
Good_Looki			122	28	206	0.136	0.250	0.239	Not
Glorious			121	35	198	0.177	0.312	0.300	40
Stunning		112	125	24	213	0.113	0.214	0.203	40
Magnificent									
Beautiful	*COCA*		107	47	202	0.233	0.440	0.377	60
Pretty	+BNC+*WIKI*+		29	6	165	0.036	0.207	0.070	50
Lovely	GloWbE		112	33	221	0.149	0.295	0.260	50
Good_Looki			122	23	241	0.095	0.188	0.174	Not
Glorious			121	35	228	0.153	0.290	0.266	50
Stunning		141	125	39	228	0.171	0.312	0.292	60
Good_Looking									
Beautiful	*COCA*		107	28	201	0.139	0.262	0.245	Not
Pretty	+BNC+*WIKI*+		29	10	141	0.071	0.345	0.132	Not
Lovely	GloWbE		112	28	206	0.136	0.250	0.239	Not
Magnificent			141	23	241	0.095	0.188	0.174	Not
Stunning			121	10	233	0.043	0.083	0.082	Not
Glorious									
Beautiful	*COCA*		107	34	194	0.175	0.318	0.302	30
Pretty	+BNC+*WIKI*+		29	6	144	0.042	0.207	0.080	Not
Lovely	GloWbE		112	35	198	0.177	0.312	0.300	40
Magnificent			141	35	228	0.153	0.290	0.266	50
Good_Looki			122	10	233	0.043	0.083	0.082	Not
Stunning		121	125	23	223	0.103	0.190	0.187	30
Stunning									
Beautiful	*COCA*		107	35	197	0.177	0.327	0.302	50
Pretty	+BNC+*WIKI*+		29	9	145	0.062	0.310	0.117	Not
Lovely	GloWbE		112	24	213	0.113	0.214	0.203	40
Magnificent			141	39	228	0.171	0.312	0.292	60
Good_Looki			122	16	231	0.069	0.131	0.129	Not
Glorious		125	121	23	227	0.103	0.190	0.187	30

Table 3.3 Comparison with UMBC Toolkit for 'beautiful'

Word Chosen: Beautiful (w)				
Synonym(x)	**Modified WebJaccard (*w,x*)**	**Modified WebOverlap (*w,x*)**	**Modified WebDice (*w,x*)**	**Toolkit(value in %age)**
Pretty	24.21	69	33.13	40
Lovely	68.61	70	77.27	80
Magnificent	42.22	54.9	53.21	60
Glorious	31.85	39.75	42.02	30
Good_Looking	25.29	32.75	34.54	X*
Stunning	32.21	40.87	42.58	50

*X means not available.

Table 3.4 Comparison with UMBC Toolkit for 'pretty'

Word Chosen: Pretty (w)				
Synonym(x)	**Modified WebJaccard (*w,x*)**	**Modified WebOverlap (*w,x*)**	**ModifiedWebDice (*w,x*)**	**Toolkit**
Beautiful	24.21	69	33.13	40
Lovely	21.66	64.62	30.03	50
Magnificent	6.55	25.87	9.87	X
Good_Looking	12.92	43.13	9.21	30
Glorious	7.64	25.87	11.28	X
Stunning	11.28	38.75	16.50	X

Table 3.5 Comparison with UMBC Toolkit for 'lovely'

Word Chosen: Lovely (w)				
Synonym(x)	**Modified WebJaccard(*w,x*)**	**Modified WebOverlap (*w,x*)**	**ModifiedWebDice (*w,x*)**	**Toolkit**
Beautiful	68.61	70	77.27	80
Pretty	21.66	64.62	30.03	50
Magnificent	27.12	36.87	36.66	50
Good_Looking	24.75	31.25	33.70	X
Glorious	32.21	39	42.30	40
Stunning	20.57	26.75	28.62	40

Table 3.6 Comparison with UMBC Toolkit for 'magnificent'

Word Chosen: Magnificent (w)				
Synonym(x)	**Modified WebJaccard(*w,x*)**	**Modified WebOverlap (*w,x*)**	**ModifiedWebDice (*w,x*)**	**Toolkit**
Beautiful	42.41	55	53.20	60
Pretty	6.55	25.87	9.87	50
Lovely	27.12	36.87	36.66	50
Good_Looking	17.29	23.5	24.53	X
Glorious	27.85	36.25	37.51	50
Stunning	31.12	39	41.17	60

Table 3.7 Comparison with UMBC Toolkit for 'good looking'

Word Chosen: Good_looking (w)				
Synonym(x)	**Modified WebJaccard(*w,x*)**	**Modified WebOverlap (*w,x*)**	**ModifiedWebDice (*w,x*)**	**Toolkit**
Beautiful	25.30	32.75	34.54	X
Pretty	12.92	43.13	18.61	X
Lovely	24.75	31.25	33.70	X
Magnificent	17.29	23.50	24.53	X
Stunning	7.83	10.3)	11.56	X

Table 3.8 Comparison with UMBC Toolkit for 'glorious'

Word Chosen: Glorious (w)				
Synonym(x)	**Modified WebJaccard(*w,x*)**	**Modified WebOverlap (*w,x*)**	**ModifiedWebDice (*w,x*)**	**Toolkit**
Beautiful	31.85	39.75	42.58	30
Pretty	7.64	25.87	11.28	X
Lovely	32.21	39	42.30	40
Magnificent	27.85	36.25	37.51	50
Good_Looking	7.83	10.37	11.56	X
Stunning	18.75	23.75	26.79	30

Table 3.9 Comparison with UMBC Toolkit for 'Stunning'

Word Chosen: Stunning(w)				
Synonym(x)	**Modified WebJaccard(*w,x*)**	**Modified WebOverlap (*w,x*)**	**ModifiedWebDice (*w,x*)**	**Toolkit**
Beautiful	32.21	40.87	42.58	50
Pretty	11.28	38.75	16.54	X

Lovely	20.57	26.75	28.62	40
Magnificent	31.12	39	41.17	60
Good_Looking	12.56	16.37	18.19	X
Glorious	18.75	23.75	26.79	30

The computed results have been utilized for generation of a Fuzzy Rule Base to be used for further applications.

3.4.1 Fuzzy Rule Base

The computed results show quite an agreement with the toolkit. The errors are generated due to various factors owing to the usual ambiguity of the natural language which justify the need for fuzzification. Also, it is hard for a user to select appropriate one out of the available values, therefore fuzzification plays a vital role. The domain of discourse for the fuzzy sets is [0,100] and the chosen fuzzy sets are:

NS: Not_Similar
PrS: Poorly_Similar
SS: Somewhat_Similar
QS: Quite_Similar
PfS: Perfectly Similar

Now, one can choose an appropriate index and use it for fuzzification. The membership graph of various fuzzy sets is shown in Fig. 3.2.

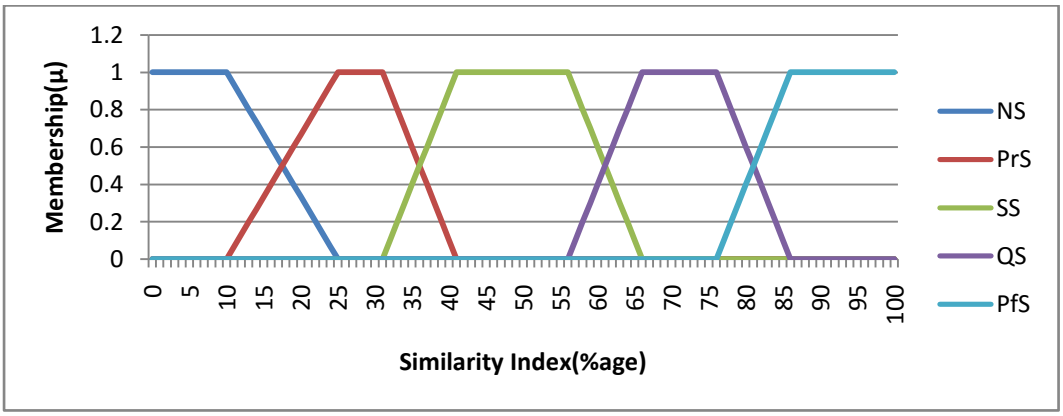


Fig. 3.2 Membership Graph of various Fuzzy sets

The design of fuzzy sets follows the Fuzzy Rule Base (FRB) needed for automated decision making. The design of fuzzy sets and the FRB is illustrative and designer of the system can modify them as per the requirement.

3.4.2 Applications of Fuzzy Rule Base

Following are the possible application areas where the proposed FRB can be applied.

- If the similarity index is *perfectly similar* use W2 in place W1 for word sense disambiguation, web page classification and query expansion.
- If the similarity index is *quite similar* use W2 in place W1 for query expansion and web page classification.
- If the similarity index is *somewhat similar* use W2 in place W1 in query expansion.
- If the similarity index is *poorly similar* use W2 in place W1 for query expansion after checking the context.
- If the similarity index is *not similar* do not use W2 in place W1.

3.4.3 Format of Next Generation Lexical Resources

The ultimate goal of the work done in this chapter is to augment the existing lexical resources by adding the similarity index and the context set in which two words can be semantically similar. The desired format of next generation lexical resources will look as follows:

Modified Lexical Resource

<Chosen Word> : <Set of possible contexts>
<Set of most commonly used contexts>
<Synonym-1>:<similarity index>
<common context set>
<Synonym-2>:<similarity index>
<common context set>
……..
……..
……..

<Synonym-k>:<similarity index>

<common context set>

The above format proposed for the lexical resources not only provides the set of synonyms but also provides the extent of similarity in the normalized manner. The similarity index can be used in the automated manner through the creation of Fuzzy Rule Base for the purpose of query rephrasing and word sense disambiguation.

The augmentation of the context set with the word enables the automated resources like web search engines in dealing with the query sense making the web search process relevant.

3.5 SUMMARY

This chapter has described an efficient and effective synonym resolution approach by finding semantic similarity between words depending upon their contexts. To avoid biasness multiple corpora have been considered for implementation. The similarity index has been computed on the basis of commonality of the contexts. Various benchmark indices have been used to find the similarity index and the results have been normalized. The results obtained have been compared with a standard toolkit for the purpose of authentication.

It may be observed that the benefit of the proposed technique is to overcome the problems of synonymy and polysemy over the information retrieval field, by finding the semantically similar words with respect to query. Moreover, an intelligent approach is used that uses the context sets to identify the words that are semantically similar. This approach also helps in enrichment of lexical resources.

The next chapter introduces a new method to find the synonyms when an entity is used as a query component instead of the attribute.

CHAPTER IV
DYNAMIC ENTITY RESOLUTION

4.1 INTRODUCTION

In the first chapter of the thesis, we described that a query contains four basic components: *keywords*, *attributes, entity* and *concept*. In the previous chapter the attribute component of the query was considered and it was explained how attributes can be replaced by their appropriate synonyms for the purpose of query expansion, resolution and reformulation. In this chapter, the entity part of the query has been taken care of and it is described how web can be used to create meaningful synonyms for a given entity to facilitate query expansion, resolution and reformulation. Before taking up the entity resolution process, let us take up the basic terminologies associated with the entities.

4.2 BASIC TERMINOLOGIES

Entity: An entity refers to a place, person, thing, event or abstraction having a distinct and separate existence from other instances of similar attributes. The reference to the entity may be local or global depending upon the context of the underlying domain.

Entity Identifier: Formal nomenclature for the entity e.g. *The Times of India*, *The Hindustan Times, Kabhi-Kabhi, Dilwale Dulhaniya Le Jayenge, i20, Santro Xing*, etc.

Entity Synonyms: A list of formal and informal identifiers referring to the same entity i.e. commonly used alternative name references to describe the entity under consideration e.g. *TOI* and *Times of India* refer to the same entity. In the same way, *Tere Bin Laden-2* and *Tere Bin Laden dead or alive* are not different.

To mathematically define the concept of entity synonyms, consider the following assumptions:

S: Universal set of strings over an alphabet

E: Universal set of entities

E_X: A list of entities over the domain X for example E_{Movies} will be a set of entities over the movie domain.

Now we can define a function F having two arguments, the first one being an arbitrary string s ∈ S and the second one being the entity domain E_X. Then the function F(s, E_X)→ e where E_X maps the string s to a single entity e or a set of entities in the global domain E which is a superset of E_x , thereby making a local reference as global.

Entity Synonym: Two strings s1 and s2 defined over the set S are said to be entity synonyms iff F(s1, E_X)= F(s2, E_X)

Entity Hyponym: A string s1 is a entity hyponym of the string s2 (both defined over the set S) iff F(s1, E_X) ⊆ F(s2, E_X).

Entity Hypernym: A string s1 is a entity hypernym of the string s2 (both defined over the set S) iff F(s1, E_X) ⊇ F(s2, E_X).

The problem of finding the entity synonyms of a string s can be mathematically described as a situation to create a set Ws of strings w's such as:

Ws = {w ϵ S | F(s, E_X) = F (w, E_X)}

The set W = {w1, w2……wk} contains entity synonyms for the string s over the domain X. Given a string s over the domain X, we have to find out W in the context of E_X.

4.3 EXISTING WORK AND THEIR DRAWBACKS

The work discussed in the literature survey forms the basis for objectives of the proposed work with following set of identified problems:

1. Synonym sets generated through existing methods are not rich and global. They are unable to take into account the massive and heterogeneous content of the web.
2. Candidate synonyms are not generated by considering the contexts.

3. In many cases, the output is limited to only those synonyms which are substrings of the entity name under consideration.
4. In many cases, availability of candidate reference is a priori requirement which is not desirable.
5. Some existing approaches only consider relationship between titles, so they suffer from the limitation of *title only concept* without taking into the account the page content as a whole.
6. There is no method for defining an index to assess the quality of synonyms generated.
7. Most of the approaches fail to take up the synonyms for general purpose common entities.

4.4 SIGNIFICANCE OF THE PROPOSED APPROACH

The solution to the above listed challenges is to design a novel mechanism to generate quite rich and credible set of entity synonyms that can be help the search engine to refer to an the entity under consideration in different ways. Entity synonyms generated through the proposed method have an edge over prevailing mechanisms, as it provides:

- More relevant set of entity synonyms (both in terms of quantity and quality)
- An index to access the quality of generated entity synonyms.
- Fuzzification of the Index for the purpose of automation.

The work will contribute to web-search in following ways:

- Improved search relevance
- Improved user experience
- Query auto suggestion
- Creation of entity dictionary
- Meaningful query expansion for the queries involving entities.

4.5 PROPOSED APPROACH

This section presents a method for discovering entity synonyms, with application to the Web Search. The proposed work comprises the iterative utilization of Search Engine Result Pages (SERPs), extraction of context from the URL, extraction of anchor text, webpage titles & snippets and candidate synonyms from query log. The proposed approach generates the set of entity synonyms using static and dynamic data. For the purpose of static data, web query log is used as an offline source. For dynamic data, online web content is used. A fragment of the query log is shown in Appendix D.

The procedure starts with the issuance of query (an entity) by the web client on the search engine interface. The search engine gets the query and returns the result pages referred to as SERPs. In the proposed approach, the Universal Resource Locators (URLs) of these SERPs are looked in the query log to get the candidate synonyms. The title and snippets of the URLs of these SERPs are utilized to obtain the contexts. By this method, first level of candidate entity synonyms are obtained. Now, these initial set of candidate synonyms are combined with contexts in order to explore more entity synonyms using dynamic web data.

A new query is then issued to the search interface using a combination of a candidate synonyms and the context related to an entity to obtain a new set of SERPs.

For extracting rich and more focused entity synonyms, the dynamic web content is used. For this purpose, the algorithm based on *Inbound Anchor Text* is applied. The anchor text that is the clickable text in a hyperlink and is relevant to the page a user is looking for, rather than generic text. The process actually begins whenever a new query which is obtained from static method is entered onto the search interface. The search interface returns a list of URLs. These URLs are collected to form a list of parent URLs (PUs).

Next, these PUs are further treated as input to generate sub parent URLs (SPUs). Thereafter, SPUs are visited one by one and downloaded web pages are retrieved in form of child documents. All pairs (anchor text, link) contained in child documents are collected in a hash map of anchor text and its corresponding URL as a set of child

URLs. The child URLs contained in the hash map is compared with the parent URLs. If there is match between child URL and parent URL, then anchor text corresponding to child URL will act as a candidate entity synonym. The child documents are also used to find the context for input entity.

The context used by the algorithm is also retrieved using title and snippet of child documents. Context obtained are combined with the original entity (query string) to produce another set of candidate entity synonyms. Sub parent URLs are also compared with parent URLs, if match occurs then the trailing part of sub parent URL will act as a candidate entity synonyms.

Thus, entity synonym is extracted using four things:

- The user history log database
- Child map in case of match
- Trailing part of sub parent URL
- Combination of query and context obtained from child documents

The detailed process is shown through a flowchart as shown in Fig. 4.1. The snapshots of the implementation results using various techniques including the proposed one is shown in Appendix B.

4.5.1 Generation of Similarity Index

After getting the candidate synonyms, similarity index is computed between the actual entity word and the candidate synonyms using *Web Jaccard* [142] method. The index values obtained thereof are normalized between the range [0, 1]. Taking the normalized fuzzy value as the outline criteria, fuzzy sets are defined to express the quality of synonyms linguistically. These fuzzy sets are then used in Fuzzy Rule Base (FRB) for the automated application of entity synonyms in the web search process. Fig. 4.1 demonstrates the strategy to generate optimized set of entity synonyms.

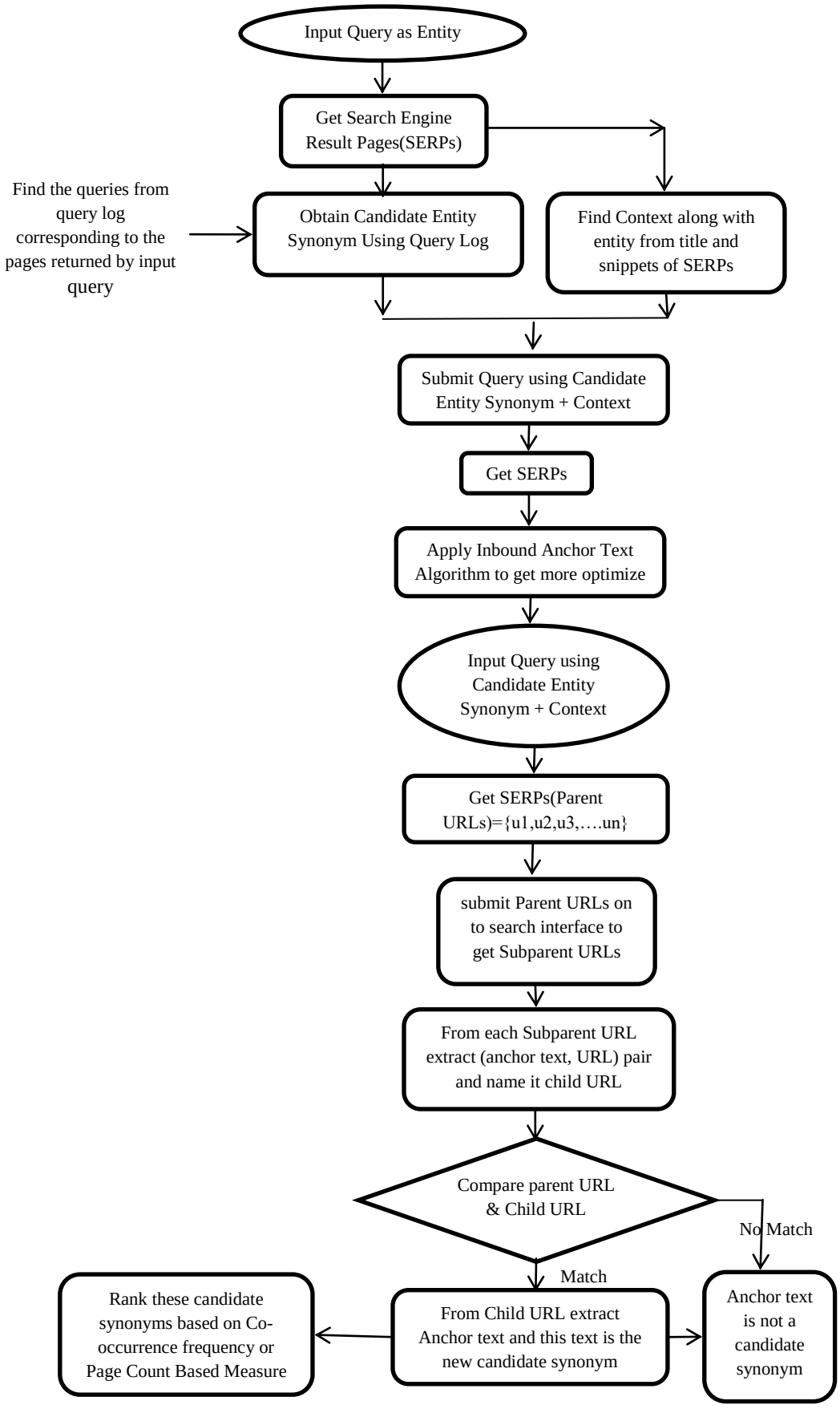


Fig. 4.1 Strategy to Generate Optimize Entity Synonyms

4.5.2 Dataset Description

Various online resources and web data (online) have been utilized for the purpose work, the details of which are given as under:

(i) Query Log

Query logs are used for improving user search experience because they act as a resource to explore the history of the user for a specific time period. As a dataset, AOL[143] query log for the period of one year is used. This log consists of more than 40 million entries. The structure of the query log is shown in Table 4.1.

Table 4.1 Structure of Query Log

Anon ID	Query	Query Time	Item Rank	Click URL
5383757	times of india	2006-03-13 12:36:09	1	http://timesofindia.indiatimes.com
5175703	Near death experiences	2006-05-09 22:36:08	2	http://www.nderf.org
470385	george bush	2006-03-30 15:44:00	3	http://bushlibrary.tamu.edu

The data set includes (AnonID, Query, QueryTime, ClickedRank, DestinationDomainURL), whose descriptions are given below:

- AnonID: An anonymous user ID number.
- Query: The query issued by the user
- QueryTime: The time at which the query was submitted for search.
- ItemRank: If the user clicked on a search result, the rank of the item on which they clicked is listed.
- ClickURL: If the user clicked on a search result, the domain portion of the URL in the clicked result is listed.

Each line in the above table represents one of two types of events:

1. A query that was NOT followed by the user clicking on a result item.
2. A click through on an item in the result list returned from a query.

Here, the ClickURL attribute in the table get matched with the Search Engine Result Pages (SERPs) to obtain the basic set of candidate synonyms. The very large file (query log) is used after splitting into a number of sub files, and then parallel processing is used to scan these files to get the desired results.

(ii) Web Data

Web data is the data that comes from large or miscellaneous number of sources. Web data are developed with the help of Semantic Web tools such as Resource Description Framework (RDF), Web Ontology Language (OWL), and SPARQL (Simple Protocol and RDF Query Language). Web content is textual, visual, or aural in nature and can be encountered as part of the user experience on websites. As a part of dynamic data, web data is used.

4.5.3 Implementation Details of the Proposed Approach

Fig. 4.2 describes the basic methodology of the proposed work and Fig. 4.3, Fig. 4.4 and Fig. 4.5 shows the architecture for different component of the proposed work.

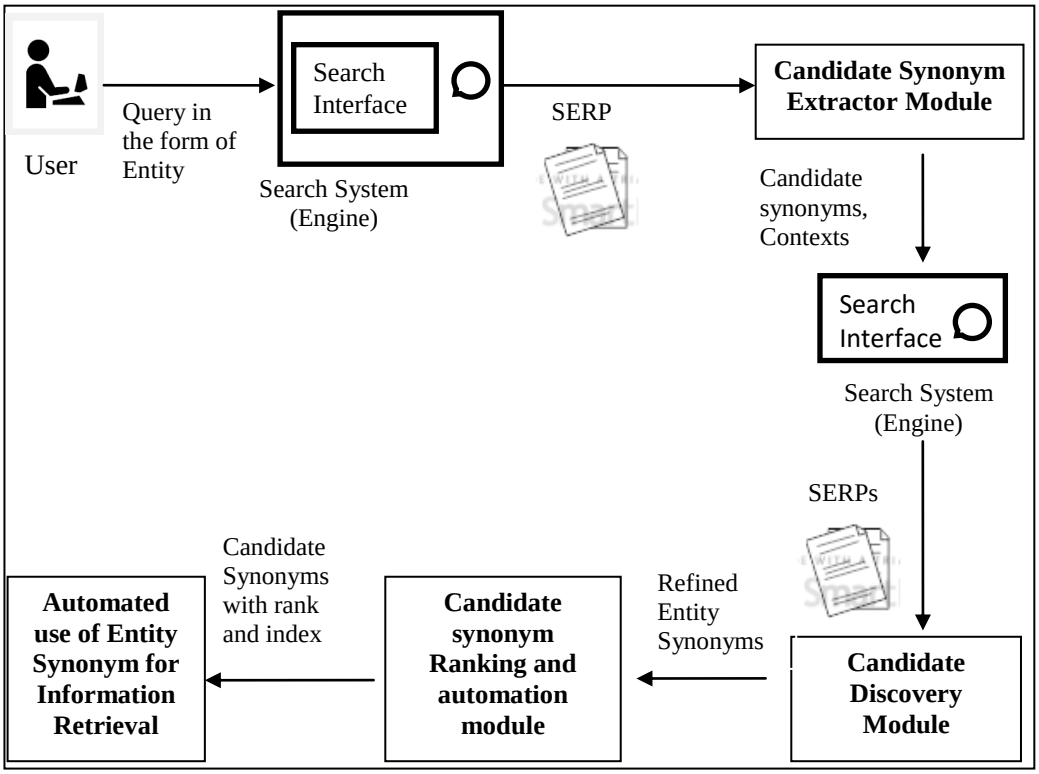


Fig. 4.2 Basic Architecture of Proposed Methodology

The purpose of different modules shown in Fig. 4.2 is as follows:

(i) Candidate Synonym Extractor Module

- This module matches the URLs returned on search interface with the URLs present in Query Log for obtaining the basic set of candidate synonyms
- It also finds the context from snippets and titles.
- It combines the basic set of candidate synonyms with the contexts to extract all possible combinations of candidate synonyms.

The detailed process of candidate synonym extractor module is shown in Fig. 4.3.

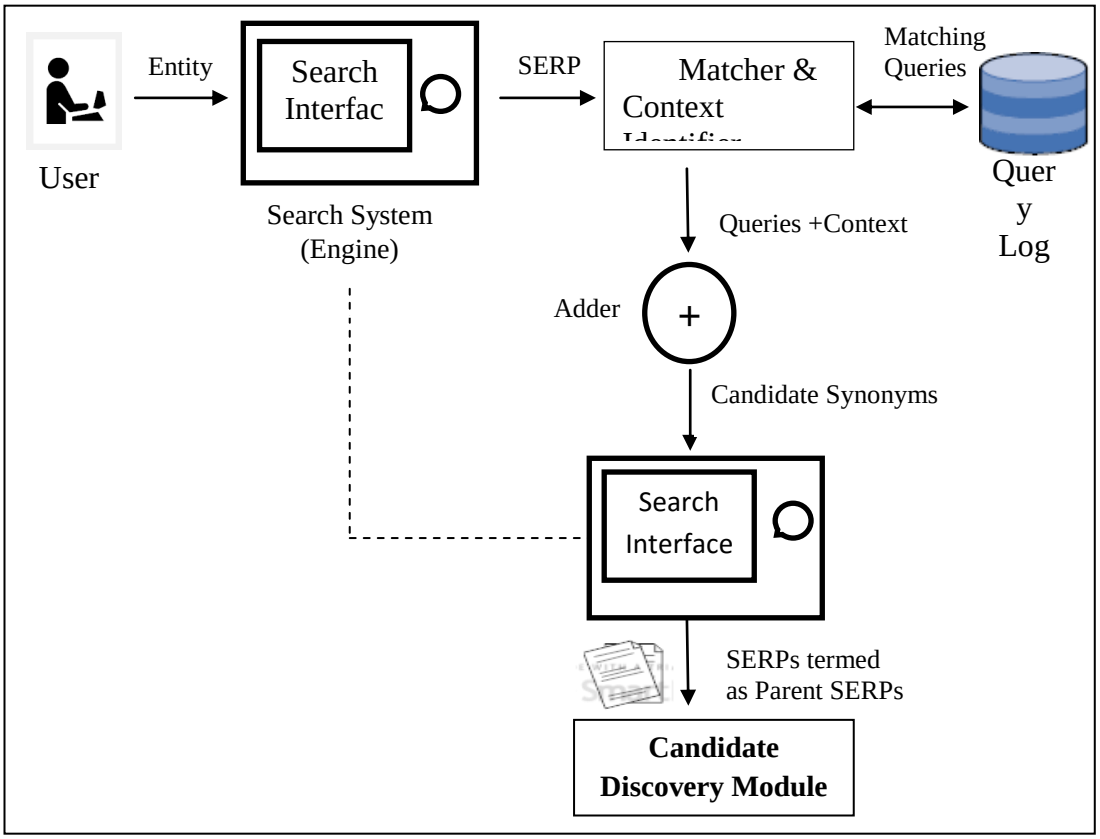


Fig. 4.3 Architecture of Candidate Synonym Extractor Module

(ii) Candidate Discovery Module

- This module finds the SubParent URLs after issuing Parent URLs on browser.
- It downloads the SubParent URLs to get the child pages.
- It extracts <anchor text, URL> pair to obtain child URLs.

- It matches the Parent URLs with Child URLs and obtains candidate synonyms as an anchor text.
- From downloaded child pages, it also extracts context from snippets and title. Contexts are also obtained from trailing part of Parent URLs.
- It combines all candidates generated from this module to get the refined set of candidate synonyms.

The process of entity candidate discovery module is depicted in Fig. 4.4.

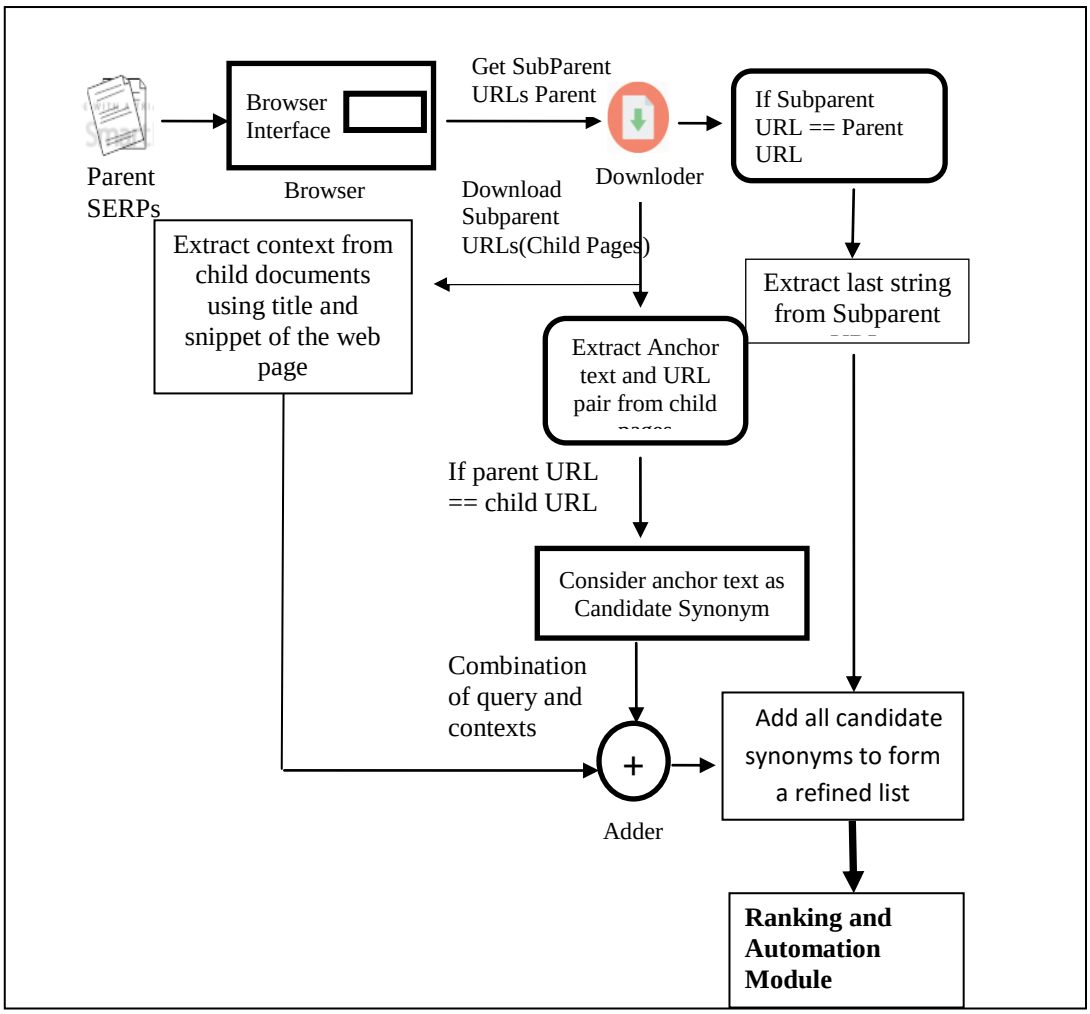


Fig. 4.4 Entity Candidate Discovery Module

(iii) Candidate Synonym Ranking and Automation Module

- This module calculates the page count for entity, page count for refined set of candidate synonyms, page count for entity and

refined set of candidate synonyms, page count by combining entity or refined set of candidate synonyms,

- It then applies WebJaccard measure to find the similarity index.
- It also applies normalization and ordering to obtain normalized and sorted index.
- Then fuzzification is done to obtain the fuzzy set.

The process of candidate synonym ranking and automation module is depicted in Fig. 4.5.

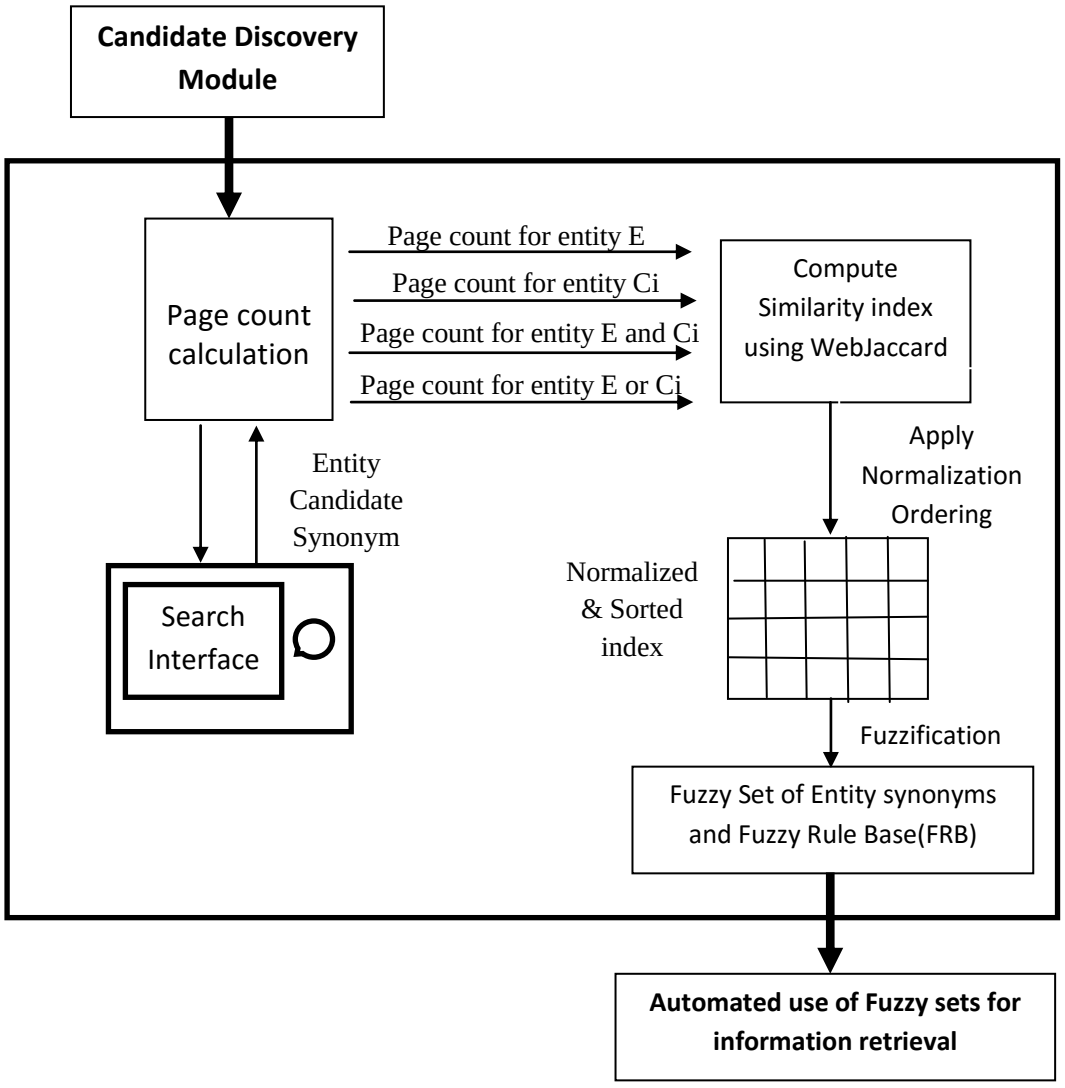


Fig. 4.5 Candidate Synonym Ranking and Automation module

(iv) Automated use of Fuzzy set for Information Retrieval

- This module helps in automated search process by the search engine using the techniques like Fuzzy Rule Base, Knowledge Graph etc.

4.5.4 Algorithm for Entity Synonym Discovery

After depicting the process, the algorithm to implement the process is outlined below:

(a) Algorithm: EntitySynonymExtractor(E,QL)

Input: Entity E, Query Log Database (QL)

Output: Ranked list of Entity Synonyms

//Algorithm to find entity synonyms corresponding to input Entity word E

1. SERPs = SearchEngine (E) //Submit Query E to interface & extract first 20 pages
2. For each p ε SERPs
   2.1 Pick the candidate entities if the URL returned by search engine is same as the URL already present in the query log. //Let it be {E_1, E_2, E_3, ……., E_n}
   2.2 Extract the URL of page p in order to retrieve context from it.
   2.3 Find context using snippets and title related to p. //Let it be {C_1, C_2 C_3…C_n}
   Make combinations of input entity string E and candidate entities obtained from query log with context obtained in step 2.3. // i.e. E_iC_i {E_1C_1, E_2C_2……E_2C_1, E_2C_2……E_nC_n} considered as entity synonyms.
3. Submit each candidate synonym as a query to the search interface
4. newSERPs = SearchEngine(EiCi)
5. Refined_Entity_Synonyms=CandidateDiscovery(newSERPs)
6. Return(Refined_Entity_Synonyms)
7. End.

Algorihm : CandidateDiscovery(newSERPs)

// Entity Candidate Discovery Module

Input: newSERPs (search result pages returned by entity_synonym_extractor)

Output: Refined_Candidate_synonym

1. Treat URLs corresponding to new SERPs as Parent URLs denoted as PU's
2. For Each URL $U_i \in PU_i$ Submit U_i on browser interface and pick 10 sub URLs corresponding to U //call the sub URLs as SPUs

2.1 For (Each SPU_i), download page corresponding to URL and name it as Child Document

2.1.1 Extract contexts from each child document

2.1.2 Extract anchor text and URL pair from each child document and store it in a map with attributes as child URL and its corresponding anchor text.

2.2 If (PU_i == child URL)

2.2.1 Extract anchor text and combine them to the list of candidate synonyms (i.e. CS_i).

2.3 If (SPUs == PU_i)

2.3.1 Extract last string from Subparent URL and add it to the list of candidate synonyms (CS_j).

3. Make combinations of input entity query string with the contexts obtained in step 2.2.1 and add them to the list of candidate synonyms (CS_k).
4. Now, union all candidate lists obtained in above steps.
   Refined_CS = $CS_i \cup CS_j \cup CS_k$
5. //Ranking Candidate Synonyms
   Retrieve the page counts for *E i.e. NE*
   For each C*i* ∈ Refined_CS do
   Retrieve the page counts for *E* and C*i i.e. NEci*
   Retrieve the page counts for C*i i.e. Nci*
   Compute WebJaccard*(E,Ci) = NEci /(NE + NCi− NEci)*
   End for

4.6 APPLICATION OF RESULTS

The proposed work is implemented by considering more than 30 user queries randomly selected from the search log of a general web search engine. The algorithm is implemented on Intel Core Duo Processor with 3 GB RAM. Software requirements include installation and setting up the environment for the software like eclipse Java neon and JDK8. Table 4.2 shows the candidate synonyms generated through different approaches including the proposed one. Some example strings are considered for comparison between various approaches. It can be clearly seen that the result set in the proposed approach is much richer than the others. Also Similarity Index (SI) has

been shown with entity synonyms computed on the basis of Web Jaccard coefficient as described in the algorithm.

Table 4.2 Comparison between Conventional and the Proposed Approach

Entity string	**Candidate synonyms generated through Query Log approach as used in [12] with SI**	**Candidate synonyms generated through Inbound Anchor Text Approach as used in [15] with SI**	**Contexts Extracted from Title & snippets**	**Candidate synonyms generated through Anchor Text+ Context with SI**	**Candidate synonyms generated through proposed Methodology (Hybrid of Static & dynamic approaches combined with context) with SI**
near death experience	1. edgar cayce heart=0.152 2. predictions edgar=0.088 3. fear of death=0.069 4. death=0.048 5. the after life=0.046 6. george anderson=0.025 7. edgar cayce on the origin of man=0.015 8. cayce on the origin of the soul=0.013	1. A site with many NDE accounts, and with some statistical analysis=0.028	• Book • Opportunity • Eben • week-long • while • Aiden • Miller • alexander • Light • walking	1. 10-astonishing-near-death-experiences=0.376 2. near-death_experience=0.168 3. A site with many NDE accounts, and with some statistical analysis=0.029	1. fear of death=0.449 2. death=0.399 3. afterlife=0.378 4. 10-astonishing-near-death-experiences=0.347 5. life-after-death=0.273 6. Near-Death Experiences and the Afterlife=0.193 7. near-death_experience=0.159 8. death_anxiety=0.153 9. overcome-the-fear-of-losing-a-loved-one=0.050 10. A site with many NDE accounts, and with some statistical analysis=0.027 11. overcome-phobia=0.025 12. overcome-fear-of-disease=0.019 13. life-beyond-death-the-science-of-the-afterlife-2=0.013 14. overcome-the-fear-of-death=0.013
animal planet	1. animal planet=0.22 2. www.animal planet.com =0.181	1. Animal Planet Live=0.505	• Tv • Planet • Twitter • Mania • Adoption • Animal • planet	1. Animal Planet Live=0.505 2. animalplanettv=0.035	1. ANIMAL PLANET - Surprisingly Human.=0.851 2. Animal Planet=0.635 3. animal-planet=0.624 4. animalplanettv=0.242 5. adoption-agencies-organizations=0.234 6. Wild Animals=0.182 7. animalplanettv=0.171 8. tv-shows=0.038 9. Animal Planet Live=0.027 10. meetanimals=0.012
indiatimes	1. times of india=0.160 2. indiannews.com=0.083 3. timesofindia=0.020	1. Indiatimes=0.932 2. Times of India=0.159	• View • India • Shopping • network	1. Indiatimes=0.932 2. indiatimescom=0.765 3. Times of India=0.159	1. the_times_group=0.977 2. list_of_newspapers_in_india_by_circulation=0.575 3. the_times_of_india=0.542 4. The Economic Times=0.400 5. list_of_newspapers_in_india_by_readership=0.283 6. toi-editorials=0.152 7. Indiatimes=0.111 8. indiatimesshopping-coupons=0.078 9. Times View=0.077 10. hindustan_times=0.065 11. times-views=0.050 12. times-news-network=0.038 13. TOI Edit=0.036 14. list_of_newspapers_in_india=0.026

					15. Times of India=0.017 16. the_economic_times=0.015 17. ePaper=0.012
superpages	1. yellow pages=0.020	1.Superpages.com=0.321	• City • One • Australian • Representation • superpagescom	1. Superpages About Page=0.973	2. Yellow pages=1.000 3. About Whitepages Pro=0.971 4. Whitepages Pro=0.883 5. Superpages About Page=0.448 6. whitepages=0.217 7. white-pages=0.195 8. Back to Whitepages=0.119 9. australian-business-directories-local-seo=0.015 10. www.whitepages.com.lb=0.103 11. yellow_pages=0.100 12. superpages-rev=0.040
newton's law of motion	1. newton laws of motion=0.011		• mathematician • physicist	1. newton-s-laws=0.090	1. newtons-laws=0.374 2. newtons-laws-of-motion=0.287 3. newtons-second-law-formula=0.062 4. physics-tutorial=0.060 5. newton-s-first-law=0.052 6. newtons-third-law-motion=0.034 7. newton-s-laws=0.023 8. newton-s-third-law=0.013 9. newton039s-three-laws-of-motion=0.011 10. newton-s-second-law=0.010
david letterman	1. david letterman show=1.000	1. David Letterman=1.000	• Official • Tv • Letterman • Letterman • induct, special • michael • twitte • guest • tenure	1. David Letterman=1.000 2. letterman=0.498	1. late_night_with_david_letterman=0.931 2. davidletterman=0.735 3. late_show_with_david_letterman=0.482 4. stephenathome=0.126 5. the_david_letterman_show=0.041 6. ed_sullivan_theater=0.021 7. lateshowwithdavidletterman=0.014 8. david_letterman=0.010
walt disney world	1. disney world=0.268	1. Theme Park Tickets=0.158 2. Resort Hotels=0.086 3. See All Walt Disney Resort Destinations =0.052	• Walt\ • Fl • Resor • world	1. Theme Park Tickets=0.158 2. Resort Hotels=0.086 3. Magic Kingdom Park=0.084 4. See All Walt Disney Resort Destinations=0.052 5. resorts=0.016 6. destinations=0.015 7. attractions=0.012	1. walt_disney_world=0.724 2. resort-hotel-list=0.632 3. disney-dining-plan=0.634 4. Disney Resort hotels=0.631 5. wandering-reindeer=0.520 6. epcot-international-food-and-wine-festival=0.481 7. walt_disney_world=0.464 8. contemporary-resort=0.416 9. Epcot International Food & Wine Festival=0.215 10. epcot=0.205 11. Magic Kingdom Park=0.138 12. Resort Hotels=0.129 13. View all Dining Plans questions.=0.113 14. magic-kingdom=0.101 15. caribbean-beach-resort=0.088 16. all-star-sports-resort=0.077 17. guests-with-disabilities=0.074 18. Disney Resort hotels=0.060 19. disney-hotels-resorts=0.053 20. magic_kingdom=0.031 21. blizzard-beach=0.023 22. disneyland=0.021 23. Disney Resort hotels=0.017
stock market	1. stock exchange=0.033 2. bombay	1. Markets=0.056	• News • Trading • Game	1. stock-market=0.882 2. investing=0.088 3. markets=0.056	1. nasdaq=0.997 2. bombay_stock_exchange=0.524 3. bse-stock-exchange=0.453

	stock exchange=0.025		• Index • India • Market • market • definition • page • stock		4. sensex_30_companies=0.416 5. NASDAQ website=0.378 6. Bombay Stock Exchange=0.328 7. stocksmarketsindia=0.303 8. bombay-stock-exchange=0.232 9. national_stock_exchange_of_india=0.216 10. bse=0.122 11. stock_market=0.114 12. london_stock_exchange=0.107 13. stock_exchange=0.107 14. bse-sensex=0.086 15. domestic-index-bse_sensex=0.082 16. domestic-market-indices_bse=0.072 17. bse_sensex=0.057 18. BSE Sensex=0.056 19. BSE CD=0.046 20. hong_kong_stock_exchange=0.045 21. shanghai_stock_exchange=0.040 22. The Economic Times=0.029 23. BSEFMC=0.027 24. ET NOW=0.024 25. domestic-index-bse_bse-cd=0.017 26. et-now-live=0.016 27. sensex-live=0.014 28. capital-market=0.014 29. Live Sensex=0.014 30. SENSEX=0.012 31. shenzhen_stock_exchange=0.011
westchester	1. westchester county=0.320 2. houses built in the 1920s=0.101 3. places in westchester county to have a kids birthday party=0.018	1. Westchester County Government=0.018	• Country • Campus • Count • Program • Scene • Community • Library • Hotel • Pace • newyorkpresbyterian/Westchester	1. Westchester County Government=0.318 2. weastchester=0.076 3. los=0.032 4. county=0.023 5. village=0.022 6. thrilled=0.017 7. florida=0.013 8. illinois=0.011	1. prison-visits=0.619 2. westchester-il-us=0.556 3. westchester-map=0.426 4. rules-regulations-title-47=0.337 5. find-prison=0.314 6. radio-frequency-safety-0=0.295 7. find-prisoner=0.225 8. Westchester Library System=0.192 9. wcplnews=0.085 10. Westchester Community College=0.060 11. newyork-presbyterian-westchester-division=0.047 12. westchester-community-college-valhalla-main-campus=0.025 13. valhalla-main-campus=0.023 14. contact-westchester-division=0.023 15. westchester-division=0.021 16. Westchester County Archives=0.019
theatrehistory	1. waiting_for_godot=0.886 2. medea=0.213 3. list of greek tragedies=0.102 4. ben jonson=0.031		• Theatre • full-length	1. Theatrehistoryc om=0.054	1. waiting_for_godot=0.886 2. www.theatrehistory.com/german/goethe012.html=0.714 3. aeschylus-greek-dramatist=0.288 4. list_of_awards_and_nominations_received_by_oprah_winfrey=0.213 5. oedipus_rex=0.199 6. The Medieval Drama=0.086 7. keira-knightley-the-

	5. the play of romeo and juliet=0.023 6. the phantom of the opera=0.021 7. who is sophocles=0.016				misanthrope=0.085 8. the-misanthrope-keira-knightley-theatre=0.056 9. boris-godunov-by-pushkin=0.053 10. oedipus=0.051 11. bet-anton-chekhov=0.036 12. http//wwwtheatrehistorycom/ancient/oedipus001html=0.033 13. samuel_beckett=0.023 14. boris-godunov-literary-character=0.016 15. theatrehistory-com-ancient-oedipus001-html-1199656=0.012
desert-tropical	1. euphorbia species=0.065 2. kalanchoe=0.017	1. The Differences Between Tropical Rainforests and Deserts=0.338	• Site • major, subtropical • information • climate	1. wwwdeserttropicalscom=0.471 2. The Differences Between Tropical Rainforests and Deserts=0.337	1. list_of_deserts_by_area=0.690 2. tropics=0.666 3. tropical-and-subtropical-desert-climate=0.391 4. humid_subtropical_climate=0.188 5. jasminum_polyanthum=0.147 6. snow-bush-breynia-disticha-roseo-picta=0.146 7. climatic_regions_of_india=0.138 8. semi-arid_climate=0.065 9. sansevieria=0.064 10. rhaphiolepis_indica=0.041 11. subtropics=0.026 12. photinia=0.025 13. tropical_rainforest=0.024 14. sansevieria_cylindrica=0.018 15. photinia_serratifolia=0.016 16. pistacia_atlantica=0.015 17. rhaphiolepis=0.015 18. thar_desert=0.015 19. tropical_climate=0.014 20. desert_climate=0.011 21. rhaphiolepis_umbellata=0.011
culture vulture		1. culture vulture=0.992 2. culture vulture direct=0.652	• plugin • meaning • vulture	1. culture vulture=0.992 2. culture-vulture=0.988 3. culture vulture direct=0.652	1. flying_dutchman=0.165 2. the-flying-dutchman=0.088 3. thermionic-culture-vulture=0.087 4. thermionic-culture-vulture-super-15=0.070 5. the-flying-dutchman-richard-wagner=0.042 6. culture vulture direct=0.022 7. vulture=0.010
games brigade	1. warhammer ships=0.109		• Book • Ohio • lite	1. Shooting=0.107	1. ohiogamingbrigade=0.282 2. imperial_navy=0.111 3. black_ships=0.050 4. Ork Spacecraft=0.018 5. Imperial Navy=0.017 6. Imperial Guard=0.016 7. imperial_guard=0.016 8. Spacecraft=0.013 9. Necron Spacecraft=0.013
canon usa	1. canon camera=0.066		• World • Level • Company • Magazine • customer	1. Dslr=0.022	1. canonusa=0.253 2. canon-europe-ltd=0.160 3. cameras-lenses=0.092 4. cameras-digital-slrs=0.076 5. cameras-camera-accessories=0.066 6. eos-dslr-cameras=0.037 7. Canon Cameras=0.030 8. Canon camera lenses=0.022 9. cinema-eos=0.021 10. canon-solutions-america=0.017 11. eos-dslr-cameras=0.012
horse bows	1. historic bows= 0.0	1. Horse Bows	• Grozer	1. Horse Bows (14)=0.530	1. Grozer Csaba=0.425 2. traditionalbows=0.701

	87 2. mongolian bows=0.062	(14)=0.530			3. Assyrian recurve bows=0.376 4. Grozer Assyrian recurve bow=0.364 5. grozer-assyrian-recurve-bow=0.188 6. Albion Armorers=0.185

The proposed approach can obtain more entity synonyms as compared to conventional approaches which improves the user experience to large extent as seen in Table 4.2. In order to enumerate the effectiveness of the proposed system, precision is used as a standard metric.

The precision is calculated for each approach on the basis of the number of relevant result according to user perspective out of total number relevant results returned using each approach. The users were asked to identify the number of relevant entities in the set of returned candidate entities as shown in Table 4.3 and Fig. 4.6.

While comparing with the other existing approaches as shown in Table 4.2, it can be noticed that the proposed approach can obtain more entity synonyms, which could improve the user experience to a large extent further leads to higher precision.

Table 4.3 Number of relevant result over the total number of returned results

	Entity string	**Query**	**Anchor**	**Anchor+Context**	**Proposed**
1	near death experience	6/8	1/1	3/3	14/14
2	animal planet	1/2	1/1	2/2	8/10
3	indiatimes	3/3	2/2	2/3	17/17
4	superpages	1/1	0/1	1/1	10/11
5	newton's law of motion	1/1	0/1	0/1	10/10
6	david letterman	1/1	1/1	2/2	6/8
7	waltdisney world	1/1	2/3	5/7	20/23
8	stock market	2/2	0/1	2/2	30/31
9	westchester	1/3	0/1	2/8	15/16
10	theatrehistory	6/7	0	1/1	12/15
11	culture vulture	0	2/2	3/3	6/7
12	games brigade	1/1	0	1/1	8/9
13	canon usa	1/1	0	1/1	11/11
14	horse bows	2/2	1/1	1/1	5/6
	Avg. Precision	11.40	5.52	10.98	12.28

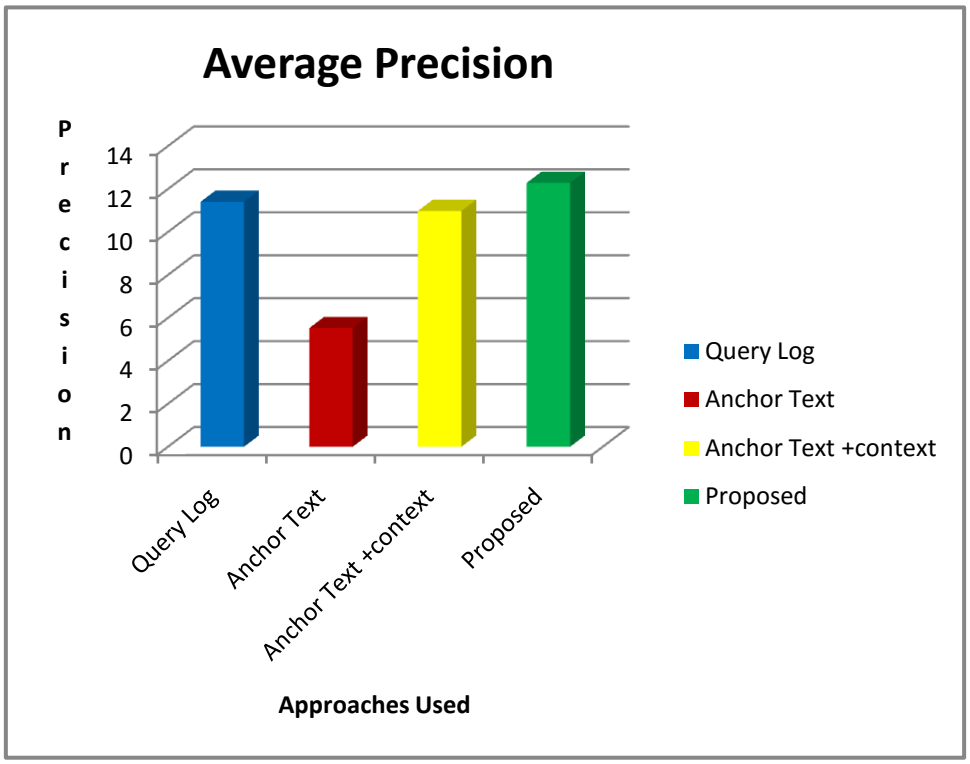


Fig. 4.6 Average Precision for Conventional and Proposed Approach

The proposed work has taken precision as the metric for evaluation of the relative performance. This metric indicates the credibility of the proposed work. We also wanted to include the metrics like Recall & F-measure in the paper, but the repository details of the big search engines and precise details about the actual number of relevant entities in their repositories is also not available. This hindered the computation of Recall and F-Measure.

The graph depicts that precision improves and the results are more meaningful in the case of proposed approch.

4.7 HALL MARK OF THE PROPOSED SCHEME

The primitive approaches use sources like Freebase and Wikipedia to generate entity synonyms for popular entities. These approaches have limited coverage and diversity in the sense that they are able to discover few or no synonyms for less popular entities.

The techniques used in past few years are based on entity source web pages and existing synonyms. Most of these approaches work on offline and structured data to

find out entity synonyms, thus, does not cater to the need of dynamic and unstructured nature of WWW.

The proposed approach combines the query log based approach, inbound anchor text and context to find the relevant and accurate candidate entity synonyms. The algorithm focuses on general query logs rather than domain-specific query logs. The query logs can be collected for a specific time frame. The contexts are identified from title and snippet of downloaded web pages which help in finding specialized resultant candidate synonyms helping the user to find better results in minimal time.

To tackle the problem of few or no synonyms for less popular entities, proposed algorithm not only uses inbound anchor text for finding candidate synonyms, but also uses snippets and title of WebPages. The algorithm also introduces a new approach for finding candidate synonyms from trailing part of sub parent URL.

When one talks of entities, it reminds him/her of *Google's Knowledge Graph*. The Knowledge Graph is Google's own database, where all of the data that has been collected from billions of wide web searches is evaluated for relevance. It is a vast graph structured representation containing so many entities and their relationships. It is also a systematic way of putting facts, people and places together and creating interconnected search results that are more accurate and relevant. We have made a small entity knowledge graph for representing entities and their candidate synonyms to the mechanism used by Google to enhance its search engine's results with semantic-search information gathered from a wide variety of sources.

The entity and its candidate synonyms are related by relationship having some similarity value. Thus, this graph can be incorporated and extended to other knowledge bases.

The Knowledge graph for the entity *indiatimes* is shown in Fig. 4.7 and small entity knowledge graph for representing entities and their candidate synonyms is shown in Fig. 4.8.

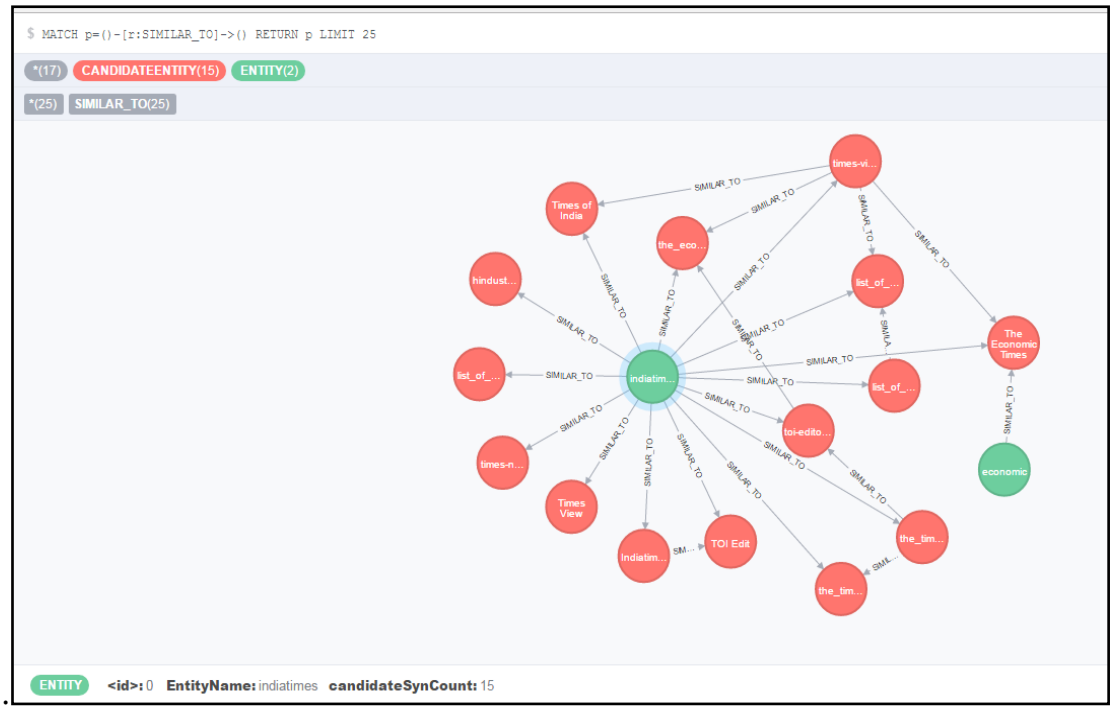


Fig. 4.7 Knowledge graph for the Entity *indiatimes*

Fig. 4.7 shows the entity and its candidate synonyms where green one represents the original entity and red node represents its candidate synonyms. The strength of relationship between the entity and its synonyms is shown through the weight of the connecting arc.

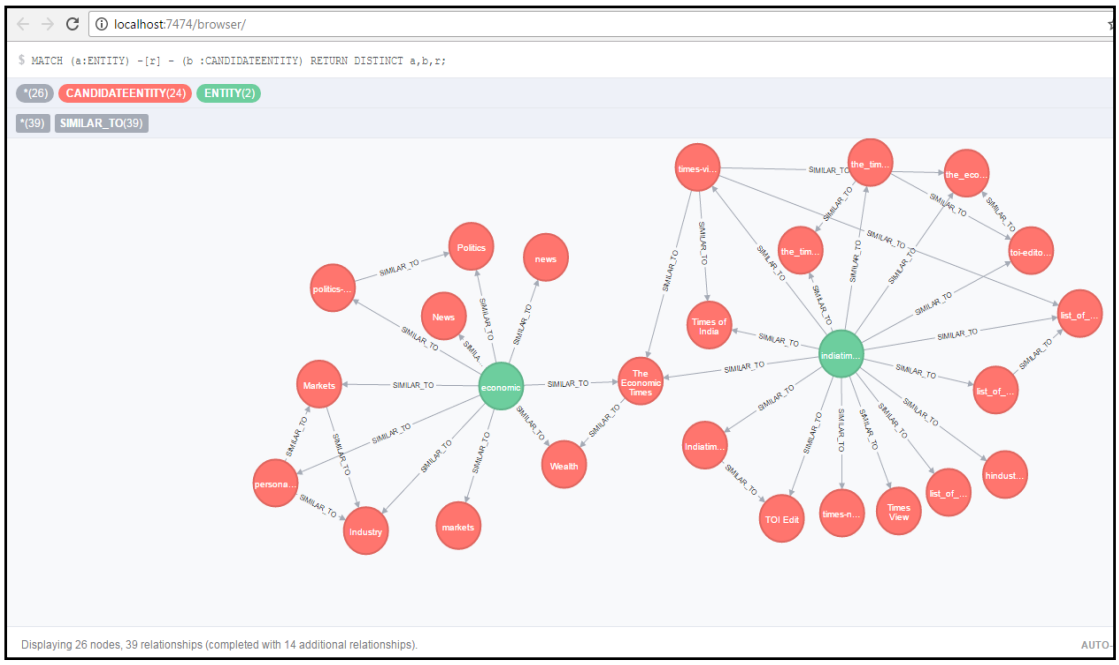


Fig. 4.8 Knowledge graph showing the relationship between two entity synonyms and their candidate synonyms

Fig. 4.8 represents the relationship between two entities and their candidate synonyms. This graph can be utilized by search engine to enhance its search results due to its ability to provide semantic search information for a variety of domains.

4.8 FUZZIFICATION OF RESULTS

The computed index results are normalized to range [0, 1]. This range is used as a domain of discourse to create fuzzy sets expressing the extent of similarity between two words. These fuzzy sets can be used for query auto suggestion, query expansion, query auto replacement etc. thereby enriching the users' search experience in the order as given in the Table 4.4.

Table 4.4 Creation of Fuzzy Sets

Name of the fuzzy set	Support	Rank	Usage
Excellent entity synonym	[0.80,1.00]	1	Auto Query suggestion, query expansion, query replacement
Very Good entity Synonym	[0.60, 0.80)	2	Auto Query suggestion and query expansion
Good entity Synonym	[0.40, 0.60)	3	Auto query suggestion
Moderate Entity Synonym	[0.20, 0.40)	4	To be used in case of poor precision
Poor Entity Synonym	[0.0, 0.20)	5	None

All the above set of generated entity synonyms can be used for various applications like:

- In information retrieval by suggesting the alternative name of the entity for getting more relevant set of documents in response to the user query.
- Creation of Entity Dictionary or Entity Knowledge Graph to enhance search.

4.9 SUMMARY

The proposed technique is scalable and can be implemented for both unstructured and dynamic Web. Moreover, it can be applied on generic as well as domain-dependent content. The results indicate that the mechanism not only provides a rich set of quality

synonyms, but also mitigates the polysemy problem to a large extent thereby providing the user with valuable and correct links. The work can be used for automated search process by the search engines using the techniques like Fuzzy Rule Base, Knowledge Graph etc.

The experimental results depict a high precision of the proposed system over other existing search systems. The approach can be used to resolve a query when it contains a named entity and returns smarter and better answers than just a matching of keywords in a query to keywords found in documents that match.

For future work, the candidate synonym set can be used for query reformulation, creation of entity dictionary for web search, named entity recognition in documents, text analytics and to extract information from unstructured data. To extend the work, more parameters can be considered to improve the quality of synonym discovery accuracy.

The next chapter presents the contributions to resolve the concepts used in the query.

CHAPTER V

CONCEPT RESOLUTION FOR FOCUSED AND ENRICHED WEB INFORMATION RETRIEVAL

5.1 INTRODUCTION

The main target of a web search engine is to understand the short piece of query text provided by the user and to give rich and pertinent information. But, the ambiguous, uncertain and inconsistent nature of natural language makes this task quite challenging.

In the previous chapters, we have taken care of *attribute* and *entity* components of the query. In this chapter, the *concept* component of the query is dealt with to direct the query towards more precise understanding of the user needs. In the case of entity and attribute, the query was made rich using the appropriate synonyms of these components. In case of concept, the corresponding query word has to be resolved through the substitution of its appropriate instance(s) using worldly knowledge making it quite a challenging task.

As discussed in the literature survey chapter of this thesis, many attempts have been made to enumerate and aggregate the worldly knowledge in the form of contributions like FreeBase [98], WordNet [97], WikiTaxonomy [112], Cyc [113], YAGO [114], KnowItAll [115], TextRunner [116], OMCS [117], NELL [118] and DBPedia [119] etc. All these contributions are manually curated and the volume of the content is quite low. The number of concepts in the WikiTaxonomy, YAGO and Cyc are between 0.1 to 0.5 million. While in FreeBase, WordNet, DBPedia and NELL, their number is in thousands. These numbers are extremely small, when one takes into account the volume of common sense knowledge associated with this world. In the practical applications requiring wordly knowledge (like web search), these resources prove to be quite inadequate. Therefore, the need has been long felt to develop the huge taxonomies and ontologies based upon the web pages in order to:

- Cover large number of concepts and their instances.

- Cover the heterogeneity and versatility of the web.
- Deal with the probabilistic or partial relationship between the concepts and entities in consideration.

All these works were started from the academic viewpoint and are unable to handle today's demand of worldly knowledge as required by the search engines.

To overcome the above mentioned problems, a project was started to gather the worldly knowledge from the web by Wentao Wu et.al. [121] in association with Microsoft. The details of the collected knowledge are available on the website of the PROBASE [145]. This knowledge contains information about the real world entities and their specific/ generic/ conceptual references with available association count. This chapter proposes and implements an algorithm for concept resolution using PROBASE. Before taking up any further details, let us have a look at the concept resolution problem.

5.2 THE CONCEPT RESOLUTION PROBLEM

When used in the query, a concept has to be substituted for its closest set of instance(s). For example, consider the queries:

Best Universities in Europe
Large Software Companies in Asia

Here, it is very hard to characterize and enlist the best universities and to find all the cities in the Europe as the phrase 'best universities' does not have a definite boundary and the word 'Europe' actually implies European cities. The worldly knowledge is too huge to be comprehended and moreover becomes ambiguous, inconsistent and uncertain at many places. Concept resolution means to make appropriate substitutions for the concepts under considerations. It requires:

- Providing the machines/systems the access to large knowledge base related to common sense vocabulary
- Enabling the machines to use this knowledge in an unambiguous manner

Both of these are challenging tasks and can't be executed to perfection. But efforts can be made to accomplish this in a quite appropriate manner. The work proposed in this chapter is an effort in this direction.

5.3 PROBASE

PROBASE is a large-scale probabilistic semantic network used for Text Understanding. It is a taxonomy that contains millions of concepts of worldly facts. Unlike traditional taxonomies that treat knowledge as black and white, it uses probabilities to model inconsistent, ambiguous and uncertain information.

The concept space employed by *PROBASE* contains millions of fine-grained, interconnected and probabilistic concepts. For each concept, a number of instances and attributes are present in the *PROBASE*. For example, a concept *company* is connected to instances such as *apple* and *Microsoft* in the knowledge base of *PROBASE*. Moreover, it also scores the concepts and instances, as well as their relationships. This abundant information allows us to build inferencing mechanisms for text analysis and text understanding. Compared with other knowledge bases such as WordNet, Wikipedia, YAGO and Nell, it has two advantages.

1. The rich concept information enables interpretation at fine levels. For example, if "China, India", are checked in the PROBASE, then it returns country and Asian country as a top concepts. Given "China, India, Brazil", the top concepts become "developing country, BRIC country, emerging market". Other knowledge bases which were used earlier do not have a fine-grained concept space, nor an inferencing mechanism for the concept, therefore they can at most map these words into the concept of 'country', which is often too general and coarse level for sophisticated text understanding.
2. Its probabilistic nature allows one to build inferencing mechanisms which map words in a context to appropriate fine-grained concepts. Moreover, it is taxonomy based upon worldly knowledge, in combination with users' statistics which results in focused and enriched outcomes. The structure of the PROBASE is shown in Table 5.1 and the sample database for the same is shown in Appendix D.

Table 5.1 Structure of PROBASE

Concept	Instance	Number of Association
Activity	Game	1871
Game	Chess	1343
Game	Poker	601
Bollywood Star	Shahrukh Khan	15

The biggest strength of the PROBASE lies in its two characteristics.

1. The taxonomy has been derived from the web, therefore it involves the actually used concepts by the people worldwide involving all sorts of heterogeneity and slang terms.
2. The size of the taxonomy is huge and contains very large number of general terms which is much bigger (by one order of magnitude) than its nearest competitors.

5.4 LIMITATIONS OF EARLIER CONTRIBUTIONS

There are many concepts that have been associated with hundred thousand instances making it quite difficult to associate the proper set of instances to the corresponding concept. Efforts in this direction include works of Wang et.al.(2012)[27], Egozi et.al.(April 2012)[21] and Sendhil et.al. (2010)[25]. These works lack depth and operate at quite a surface level. The work proposed by Wang et.al. considers short text as “Bag of Concepts” without taking into consideration the document as a whole. Explicit Semantic Analysis proposed by Egozi et.al. uses relatedness analysis based on Wikipedia but neglects the context of words and cannot exactly determine the desired sense of an ambiguous word. The work proposed by Sendhilet.al. deals with construction of personalized page view graph for small scale search which is limited to an individual only. Fonseca et.al. (2005)[22] generated and organized concept hierarchy from the stored document sets and used it for query expansion purposes with a view to improve precision. Lu et.al. (2017)[23] used TREC-VID 2015 (Multimedia Event Detection System) for handling complex concepts in the user query. Their system detected large number of concepts using pre-trained concept detectors for textual-to-visual relation. Metzler et.al.(2007)[24] proposed a new

mechanism known as ‘Latent Concept Expansion (LCE)’ for expanding the term concepts for tasks such as query suggestion and query reformulation. Boucenna et.al. (2016)[20] proposed concept-based semantic search for outsourcing the data over cloud after encrypting it.

After studying these contributions, a mechanism has been proposed in this chapter for resolving the concept(s) to its appropriate instances in the presence of available contexts such as IP address, browsing history etc. The motivating factors towards this proposal were as outlined below:

- Manually curated worldly knowledge sources such as NELL, Wikipedia, Freebase, DBPedia are insufficient to fill the requirement of the web search engines.
- There was a need to have a worldly knowledge source with the ability to handle all sort of heterogeneous and multidimensional knowledge pertaining to this world.
- PROBASE is an effort in this direction and is publically accessible on https://concept.research.microsoft.com/Home/Introduction
- Google Humming Bird principle indicates that the users’ Geographical Location, Browsing History and other such parameters can be used to cater the interest of individual user.
- In the web pages, a lot of concepts are described in the form of slang terms which are not defined in the online lexical sources e.g. biggie, bigwig, big-wig, heavyweight etc.
- Google has shifted to knowledge graph based search from keyword based search.
- A lot of time is wasted by the search engine if the same query can be interpreted in multiple manners.

5.5 THE PROPOSED CONCEPT RESOLUTION METHOD

Before moving to the proposed work, some adoptable practices are suggested which can help in concept resolution and reduce the burden on the retrieval system. These practices may include:

5.5.1 Textual Practices

Many times a query has multiple interpretations corresponding to different association of words e.g. a query *New York Times Square* can be interpreted into two ways

- New York and Times Square
- New York Times and square.

A textual protocol can be adopted wherein a hyphen can be used to resolve the ambiguity. The above two interpretations can be clarified by writing them in this fashion:

- New-York Times-Square
- New-York-Times Square

5.5.2 Concept Synonym Identification and their Merger

In PROBASE by Wu et.al.(2012)[121], numerous ideas have been distinguished independently but they can be easily merged to expand the likelihood and avoidance of uncertainty. These synonyms, not available in the lexical resources as such, can also be used for the purpose of query expansion and query recommendation. Given below are certain examples

- Celebrity, celeb
- biggie, bigwig, big-wig, heavyweight

Such slang terms though not available in the dictionary can be identified through web exploration and can be manually curated.

Concept resolution is not only helpful for the effective query recommendation, but also helps search engine to find relevant information efficiently. This work focuses on resolving the concepts from the documents to identify ambiguity of the given query, to distinguish the underlying documents based on the meaning of the query and to help web user(s) to get the desired results.

5.5.3 The Proposed Mechanism

A concept-based search query can be a collection of concepts and it can be a combination of concept and entities connected by at least one legitimate relationship(s). The instantiation process can either be direct or indirect depending upon the reality whether the concept directly resolves into the instance or resolves into the instance through a chain of sub-concepts.

It is concluded that not only the concept-concept/concept-entity relationship present in the query can be exploited as the context but even the physical parameters like IP address and the statistical data such as browsing history can also be used to estimate the intended contexts. To guarantee the delivery of intended contexts, the concept of query restructuring and backtracking has been utilized.

Fig. 5.1 demonstrates an outline of proposed concept resolution approach for the goal of web information retrieval. The sketched out system comprises of a set of modules for carrying out various functions. The Concept Identification Module (CIM) identifies the concept term(s) used in the query. This module also recognizes the possible list of entity instances associated with the concept(s) used in the query. To fulfill these tasks, CIM takes the assistance of *PROBASE*.

The extensive Concept-Entity list generated by the CIM module is submitted to the Concept Resolution Module (CRM) for settling the concept(s) to their intended entities.

The CRM gets the huge entity list from CIM related with the concept(s) used in the query. The CRM prunes this list in accordance with the contexts associated with the users' query. These contexts include geographical location with respect to IP address of the client, browsing history of the search engine etc.

The results/outcomes created by the CRM are submitted to Result Processing Module (RPM) which does the final processing, using *PROBASE*, for submission of results to the user.

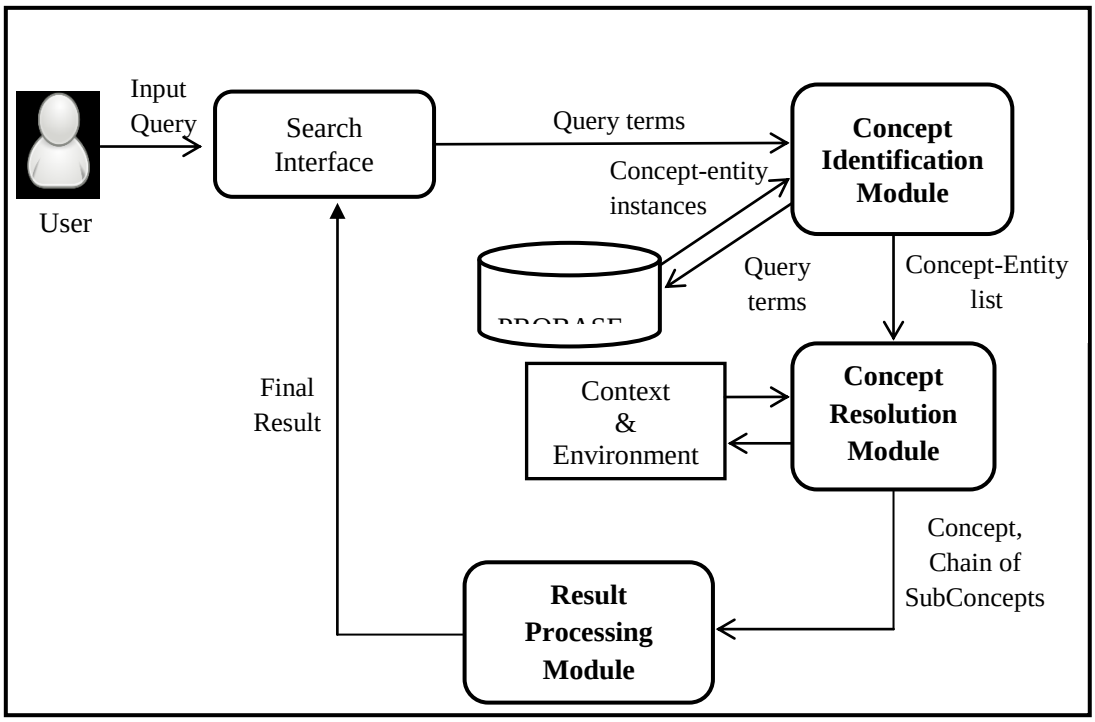


Fig. 5.1 The Proposed Concept Resolution approach for Web Information Retrieval

The working of these modules is taken up in detail in the subsequent sections.

(i) Concept Identification Module(CIM)

Fig. 5.2 demonstrates the working of Concept Identification Module. This module isolates the entities and concepts present in the input query utilizing Concept Entity Relationship File (CERF).The CERF is created by referring *PROBASE* wherein each concept present in the input query is looked for the entities corresponding to the concept are picked up. CERF is populated by isolating concept from entities using tab and all entities related to concepts using comma operator.

All entities from the input query are added to the *entity list* which can be specifically utilized by the Result Processing Module. In order to generate initial concept list, the substrings of concepts are identified that act as synonyms for the concepts present in the input query.

The concepts from the input query and their equivalent words are utilized to generate final concept list using CERF. The final concept list thus generated is passed to the concept resolution module as information.

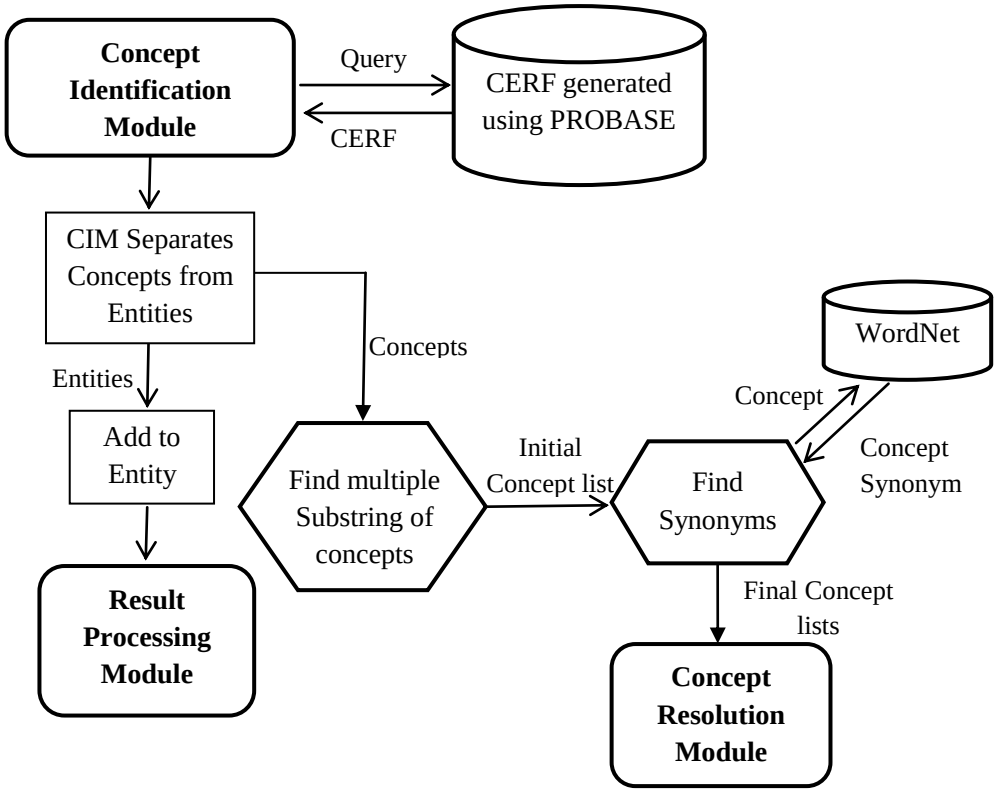


Fig. 5.2 Working of Concept Identification Module

Table 5.2 shows the sample CERF file generated from *PROBASE* which shows relationship between concept and entities simultaneously.

Table 5.2 The sample CERF file generated from PROBASE that shows relationship between concept and entities simultaneously.

Concepts	**Entities belonging to the corresponding Concept(some sample entries)**
bollywood star	Amitabh bachchan, jaya bachchan, shakti kapoor, ranbir kapoor, shahrukh khan, Katrina kaif, salman khan, shah rukh khan, priyanka chopra, shilpashetty, kareena kapoor, abhishek bachchan, aishwarya rai, john abraham, sonam kapoor, sridevi, tabu, shriya saran, vivek oberoi, ranveer singh, aamir khan, dia mirza, ompuri, preity zinta, hrithik roshan, dilip kumar, akshay kumar, sohail khan, boney kapoor, saqibsaleem, jaqueline fernandes, deepika padukone etc…
celebrity	madonna, kimkardashian, rihanna, parishilton, angelinajolie, beyonce, Jennifer lopez, lady gaga, oprah winfrey, gwynethpaltrow, britney spears, victoria beckham, jenniferaniston, mileycyrus, jessica alba, katyperry, justinbieber, brad pitt, lindsaylohan, demi moore, camerondiaz, halle berry, tiger woods, jessicasimpson, sarahjessica parker, tom cruise, davidbeckham, evalongoria, kate moss, ellendegeneres, taylor swift, leonardodicaprio, nicolerichie,

	juliaroberts, georgeclooney, kanye west, oprah, cher, johnnydepp, cindycrawford, selenagomez, pamela Anderson etc…
Entities	**Concepts belonging to the corresponding Entity(some sample entries)**
Amitabh Bachchan	bollywood star, star,celebrity, bollywood celebrity, bollywood actor, actor, prominent regional indian artiste, bollywood luminary, big name, stalwart, person, indian cinema legend, popular actor, high profile indian, famous actor, top star, bollywood celeb, biggest star, leading man, popular indian celebrity, indian cinema s stalwart, mainstream bollywood star, famous light skinned personality, personality, celeb, contemporary star, biggie, dignitary, legendary figure, eminent personality, bollywood biggie, lauded lead actor, veteran bollywood star, news report popular indian celeb, big bollywood personality, established star, film, post, great personality, iconic star, influencer, industry stalwart, modern icon, legendary actor, bollywood superstar, indian celebrity, esteemed name etc…
Salman Khan	star, actor,bollywood star, bollywood actor, bollywood celebrity, superstar, bollywood superstar, celebrity, bollywood celebrity, co-stars, biggie, a list actor, personality, film star, indian film heavyweight, popular bollywood actor, bollywood celeb, a list bollywood star, today s popular actor, visionary well funded entrepreneur, celeb, famous name, pioneer, big star, famous bollywood actor, film personality, co star, social entrepreneur, celebrity figure, successful actor, money making super star, indian actor, hindi star, bollywood personality, b town celeb, bollywood s biggest name, bollywood biggie, top bollywood star, mass heroe, leading super star, featuring keynote speaker, top bollywood actor, famous bollywood film star, bollywoods top star, meagstars, non pashto speaking pakhtoons, top star, popular bollywood superstar, luminary, popular actor, a-list actor, top actor, lead actor, bollywood's celebrity, indian mega star, top film personality, muslim actor, manyabollywood star, bollywood super-stars, top industry personality, super star etc…

(ii) Concept Resolution Module(CRM)

Fig. 5.3 demonstrates the working of Concept Resolution Module. This module refines the final concept list created through CIM using concept synonym identification and their merger, by tracking IP address, using browsing history, query restructuring and through the use of typical associations. It checks the browsing history to locate the matching concepts, which acts as a source of query suggestion on the search interface for the user to help him/her in rephrasing the query in order to get

the focused and intended results. The module then tracks the IP address of the user (based on geographic location of the user) to produce sub(sub)concept list.

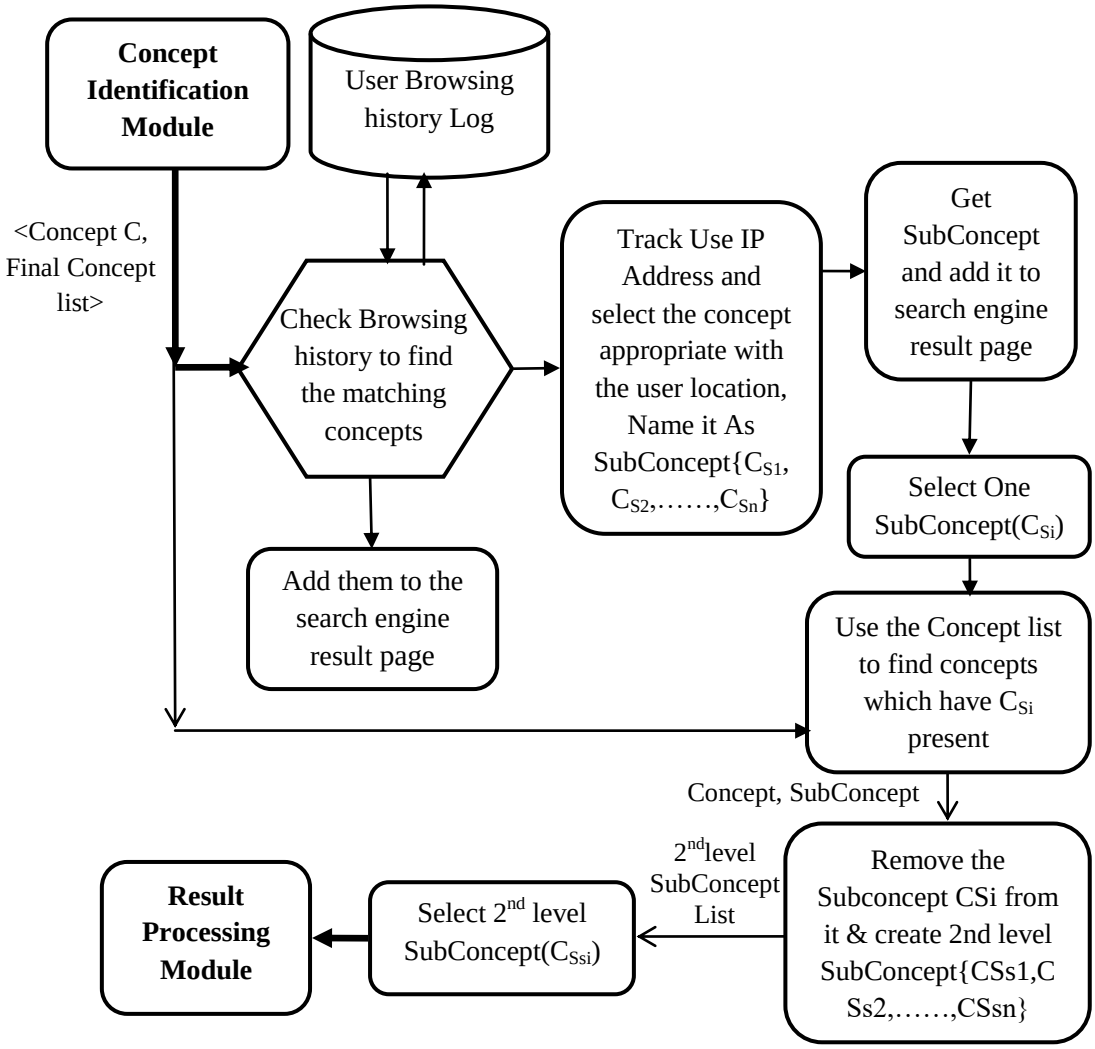


Fig. 5.3 Working of Concept Resolution Module

The motivation behind the IP address usage is to go for the localized orientation of the query because normally one looks for the information identified with his/her local domain. For example, an actor for a U.S citizen is likely to be *hollywood actor* and for a indian citizen is likely to be a *bollywood actor*. For the concept list generated so far the sub-concepts are investigated which can in turn produce new sub-concepts or the entities. The entities generated are put into search engine result pages and the sub-concept investigation continues till they are at last changed over into their associated entities. This completes the concept resolution process.

(iii) Result Processing Module (RPM)

Fig. 5.4 demonstrates the working of Result Processing Module. This module applies backtracking to distinguished entities which belong to majority of the concepts generated in the previous modules. An entity belonging to or having relationship with larger number of concepts is a probable candidate for instantiation. For this purpose ranking given in the *PROBASE* which gives the number of associations between the concept and instance on the web has likewise been utilized. The use of backtracking and number of associations can help in resolving the concept to their intended instances.

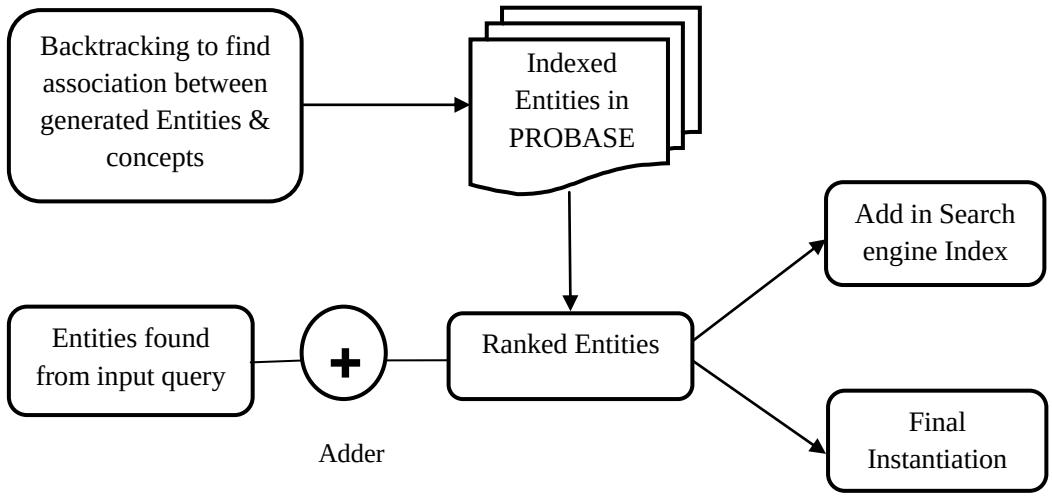


Fig. 5.4 Working of Result Processing Module

The implementation of the proposed system has been done by creating Concept Entity Relationship File (CERF) with the help of PROBASE which is further used to create the concept list through the chain process wherein at each level of the chain, sub-concepts and entities are generated and the process continues till no concept is pending for instantiation.

This chain of concepts is then backtracked for creating various query suggestions that can be used by any naïve user to get the desired results.

5.5.4 Example Illustration

The working of RPM can be explained by taking an example query say "Actor"

- Main Query : Actor
- Sub-concept retrieved at level one: Bollywood, Tollywood etc.

In order to move to the next level of concept hierarchy, user can choose any one of the sub-concepts from the available list of sub-concepts. For instance, if user chooses "*Bollywood*", it would be presented with a list of sub (sub) concepts at the next level

- Acclaimed, Established, Top, Hot etc.

At this level, user can choose any one sub (sub) concept, e.g. "Acclaimed". The RPM module backtracks the chain of concepts found at each level and restructures a new query for getting the proper instantiation for the input query. The chain of concepts can be used to obtain various query suggestions at this level.

{Acclaimed Bollwood Actor, Aamir Kahn, 5}
{Acclaimed Bollwood Actor, Amitabh Bachchan, 4}

Here first argument return the user' query, second argument indicates the instance related to the user's query and third argument indicate the number of association between the concept (first argument) and instance (second argument).

Query Suggestions are:

{Acclaimed Bollwood Actor}
{EstablihedBollwood Actor}
{Top Bollwood Actor}
{Hot Bollwood Actor} etc.

5.5.5 Algorithm for the Proposed Approach (Concept Resolution Algorithm)

The algorithm uses Concept instance knowledge base as an input and creates a concept entity relationship file which is further used to generate the entities corresponding to the concept along with an update browsing history file.

Input : Concept Instance knowledge base using PROBASE, Environmental details (IP address and the geographical location, Browsing history log)

Output : Concept Entity Relationship File (CERF), Refined set of Entities, Updated Browsing history file

Subconcept list= null

2^{nd} level Subconcept list= null

final concept list= null

Step 1. Read Concept Instance Knowledge base

Step 2. For Each distinct Concept C_i from Concept Instance Knowledge Base

2.1 Pick the entity corresponding to the concept

2.2 Add it to CERF by separating concept from entity using tab and all entities related to concepts using comma.

Step 3. Return CERF.

Step 4. Read the Concept Entity Relationship file and search for Entities present in the query

4.1 Make an arraylist "Entity" with unique entities found

Step 2. Find the substring of concept found in the input query, Name it C1

4.2 Use C1 to find more concepts from PROBASE and add it to final concept list

Step 5. Use Browsing history of the user to see any of the matching Concepts previously searched. (Take these Concepts in subconcept List).

Step 6. If (browsing history==null || no substring match occurs) then use IP address to find the related concepts from the concept list.

Step 7. Track the IP address of user according to the geographic location of user and add it refine the subconcept list.

7.1 Select one value from the subconcept list

Step 8. Corresponding to the subconcept selected, See the final Concept list to find Concepts which have the selected subconcept as a Substring.

Step 9. Select all those Subconcepts.

Step 10. Remove the subconcept string from these and Display on the Window as 2^{nd} level Subconcepts (Hyperlinks).

Step 11. Select one hyperlink.

Step 12. Apply Backtracking to find the Entities related to the concept, subconcept and 2nd levelSubconcept and print it on the window as decreasing values from the CERF file.

Step 13. Print Entities found directly in the Query.

Step 14. Update the browsing history file by adding users' past behaviour.

Fig. 5.5 shows the working of the proposed algorithm for the concept popular celebrity as an example and the working of the proposed algorithm for the concept actor is shown in Appendix C.

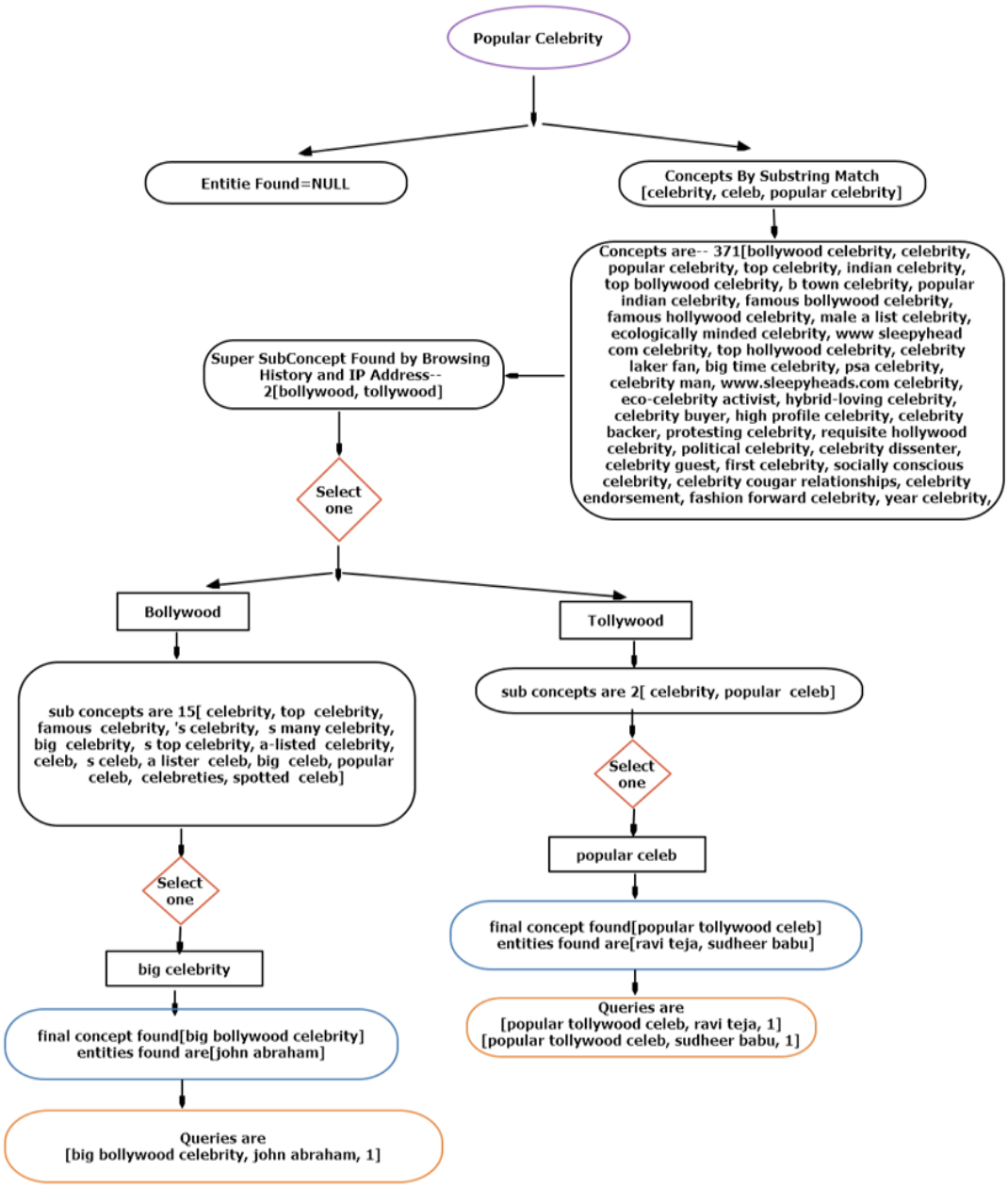


Fig. 5.5 The working of the proposed algorithm for the concept popular celebrity

5.6 IMPLEMENTATION RESULTS

The snapshots of the proposed system for the query *actor* and *bollywood actor* are shown in Appendix C. The results have been compiled using different search engines namely Google, Bing, Yahoo and the proposed system through the execution of the two sets of queries with first set containing 5 queries and second one containing 7. First experiment had 57 users and second one 43. The snapshots of the resulting pages for the two sets of queries on different search engines and the proposed system are shown in Table 5.3 & Table 5.4 and Fig. 5.6 & Fig. 5.7 respectively. In order to quantify the effectiveness of the proposed system, precision is used as a standard metric. The precision is calculated after running query at one level and also when user moves to second level in the concept hierarchy.

Table 5.3 Precision of various search engines & proposed system for one level query

Query	Bing	Yahoo	Google	Proposed
Bollywood_Actor	36.73	38.19	42.14	61.43
Hollywood_Actor	38.68	25.59	30.63	40.40
Types_of_Food	27.73	28.04	35.19	37.60
Italian_Food	30.30	39.46	40.35	45.61
Music	38.22	34.88	40.65	48.71
Disaster	40.10	36.63	32.49	48.59
Automobile	25.55	37.01	38.79	46.40
Average Precision	33.90	34.26	37.18	46.96

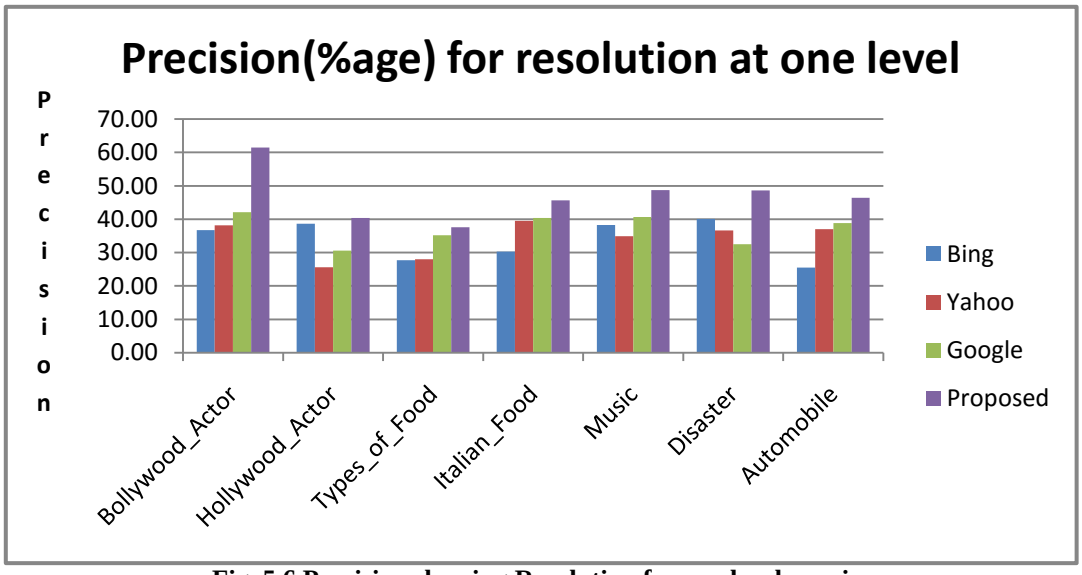


Fig. 5.6 Precision showing Resolution for one level queries

The experiment involved running a set of queries on the proposed system and the commonly available search engines. The users were asked to identify the number of relevant query suggestions made by various search engines and the proposed system. It can be observed from Fig. 5.6 and Fig. 5.7 that precision improves after exploring one more level in the concept herarchy and the results are more meaningful in the case of proposed approch i.e. precision improves when one moves towards specialized concepts in the concept hierarchy.

Table 5.4 Precision of various search engines & Proposed system for two level query

Queries	Bing	Yahoo	Google	Proposed
Actor	34.73	40.19	46.14	61.43
Food	25.73	30.04	35.19	40.60
Music	38.22	34.88	42.65	48.71
Disaster	42.10	52.63	38.49	59.59
Automobile	55.55	37.01	40.79	57.40
Average Precision	39.27	38.95	40.65	53.55

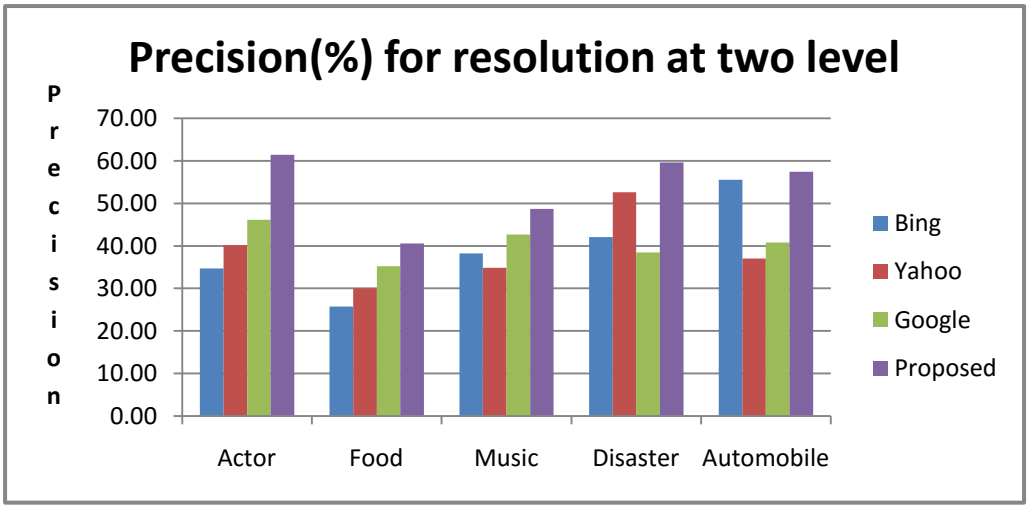


Fig. 5.7 Precision showing Resolution for two level queries

In the last phase of the system, a new query is presented to the user as a query suggestion which is having the most appropriate sub-concept corresponding to the concept present in the input query. This allows users to frame a proper search query

with the knowledge of domain terminology which will help the search engine to get the desired results.

Fig. 5.7 depicts the precision of proposed system and other search engines for various queries to measure the accuracy using traditional keyword search. It is clearly distinguishable that proposed system depicts cutting-edge results over keyword based search.

5.7 SUMMARY

The chapter presented the details about a novel approach of concept based information retrieval using concept resolution on the basis of sub-concept identification, IP address location, browsing history of the user, typical association and query rephrasing & backtracking. For the requisite worldly knowledge, the *PROBASE* has been used. This requirement has its own overhead which cannot be eradicated. The only possible alternative is to reduce the impact of this overhead on the search time by running the parallel threads to explore the worldly knowledge source in the faster manner. The working model for the same will be able to resolve the concepts to quite large extent enabling the search engine to provide the user with the meaningful and relevant results. A method is proposed for the reformulation of a query by rewriting the original query to better match the user needs such that precision and recall can be improved without affecting the original intent of the user. In fact, the experimental results depict a high precision of the proposed system over other existing search systems. The algorithm offers a mechanism which provides large scale generalization created by the voluminous worldly knowledge to the specific requirement of the user. The proposed system is simple and is able to provide ease to web users to build a proper search query with the knowledge domain terminology which will help search engine to get the desired results.

Next chapter presents the results thus obtained and their detailed discussion.

CHAPTER VI

CONCLUSION AND FUTURE ENHANCEMENT OF PROPOSED WORK

6.1 CONCLUSION

As the volume of information on the WWW continues to increase on daily basis, almost all the information available in almost all the domains is accessible on it in today's scenario. Not only the volume of information is increasing on continuous basis, but more and more heterogeneity is also becoming the part of it as the contributions are coming from around the world involving linguistic, cultural and geographical differences.

The low cost of data usage and anywhere/ anytime availability of information has proved to be a motivating factor for seeking the information from the web instead of other resources like encyclopaedia and libraries. But, the crux of the situation is:

Query is still a very short piece of text.

Exploring such a gigantic volume of information with the query text is becoming more and more challenging task for the search engines. So, there is a need to create different semantically similar versions of the query to cover the entire spectrum of the information sought. This can be done by finding out the semantically similar versions of the words used in the query (word synonyms) and similar names of entities (entity synonyms) which are compatible worldwide and have the capability to deal with the heterogeneity of the web. Also a concept used in the query must be translated to its appropriate set of instances in the light of worldly information. The work carried out in this thesis is an effort in this direction.

A summarization of the carried out work is as follows:

- The work presents an efficient and effective approach for finding semantic similarity between words depending upon their contexts. The outcome of the work includes context set identification for a word and computation of an

index indicating the extent of semantic similarity between a pair of words. The computed index has been fuzzified into a FRB for the purpose of automation which can be further utilized by the existing web search engines and other such applications. Thus, a modified lexical resource has been proposed along with several query replacement techniques in order to get efficient and quality search engine results pages.

- It proposes and implements a credible method to generate a rich set of global entity synonyms for the commonly used entities using web data wherein the availability of the candidate data is not a priori requirement. The work also proposes an index to assess the quality of entity synonyms generated which is further normalized and fuzzified for implementing automated search. It also tackles the problem of few or no synonyms for less popular entities. The proposed technique is scalable and can be implemented for both unstructured and dynamic Web.
- It enables search engines to associate the concept used in query with appropriate set of instances using a worldly knowledge source called PROBASE in the light of the factors like user's browsing history, geographical location and IP address etc.
- It also discusses the textual practices for phrase sense disambiguation for meaningful web search.
- The proposed work deals with poor quality queries by finding most appropriate replacement of original query. Thus, it helps to discover relevant information as per user query.

It is hopeful that the proposed work shall be immense help to the information and computer science professionals.

6.2 FUTURE ENHANCEMENT

The work can be further extended by devising more refined methods which are able to take up the heterogeneity of the web in the simplistic and convincing manner and construct effective rephrased queries which cover the larger spectrum of the information. Also, more and more sources of worldly knowledge sources can be

created and utilized to ensure the more effective translation of the concept to its appropriate set of instances.

Some of the possible extensions and issues that could be further explored in the near future are as follows:

- The semantic search system proposed in this dissertation can be extended to serve complex user queries, besides serving topical and informational queries.
- The work can be extended by the inclusion of more parameters and application of sophisticated techniques.
- The proposed method works on query elements. The query segmentation process has been left for the search engine and can be worked on in future.

References

[1] Tao Cheng, Hady W. Lauw and Stelios Paparizos, “Entity Synonyms for Structured Web Search”, *IEEE Transactions on Knowledge and data engineering*, Vol. 24, No. 10, pp. 1862 – 1875, October 2011.

[2] Hamid Mousavi, Shi Gao and Carlo Zaniolo, “Discovering Attribute and Entity Synonyms for Knowledge Integration and Semantic Web Search”, *In Proceedings of the 3rd International Workshop on Semantic Search Over the Web*, Riva del Garda, Italy, ISBN: 978-1-4503-2483-0, August, 2013.

[3] Surajit Chaudhuri, Venkatesh Ganti and Dong Xin, ”Mining Document Collections to Facilitate Accurate Approximate Entity Matching”, *In PVLDB*, Journal Proceeding of VLDB Endowment, Vol 2, Issue 1, pp. 395-406, August 2009.

[4] Kaushik Chakrabarti, Surajit Chaudhuri, Tao Cheng and Dong Xin, “A Framework for Robust Discovery of Entity Synonyms”, *In KDD '12 Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, Beijing, China, ISBN: 978-1-4503-1462-6, Pages 1384-1392 August 2012.

[5] Benjelloun, Omar, Hector Garcia-Molina, David Menestrina, Qi Su, Steven Euijong Whang, and Jennifer Widom, “Swoosh: A Generic Approach to Entity Resolution,” *The VLDB J.*, Vol. 18, pp. 255-276, 2009.

[6] Bhattacharya, Indrajit, and Lise Getoor, “Collective Entity Resolution in Relational Data,” *ACM Trans. Knowledge Discovery from Data(TKDD)* , Vol, 1 Issue 1, March 2007.

[7] L Jiang, Lili, Ping Luo, Jianyong Wang, Yuhong Xiong, Bingduan Lin, Min Wang, and Ning An, “An entity-relation graph based framework for discovering entity aliases” *In ICDM, IEEE 13th International Conference*, pp. 310-319, 2013.

[8] Yanen Li, Bo-June (Paul) Hsu, Cheng Xiang Zhai and Kuansan Wang, “ Mining Entity Attribute Synonyms via Compact Clustering”. *In CIKM '13 Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, San Francisco, California, USA, pp. 311-316, October 27 - November 01, 2013

[9] Roi Blanco, B. Barla Cambazoglu1, Peter Mika, and Nicolas Torzec, "Entity Recommendations in Web Search", *In International Semantic Web Conference*, ISBN: 978-3-642-41337-7, pp. 33-48, October 2013.

[10] Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka, "A Web Search Engine-Based Approach to Measure Semantic Similarity between Words", *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, NO. 7, pp. 977 – 990, July 2011

[11] Li, Yuhua, Zuhair A. Bandar, and David McLean, "An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources", *IEEE Trans. Knowledge and Data Eng.*, Vol. 15, No. 4, pp. 871-882, July-Aug. 2003.

[12] Adhikesavan, Kavitha, "An Integrated Approach for Measuring Semantic Similarity between Words and Sentences using Web Search Engine", *International Arab Journal of Information Technology, IAJIT*, Vol 12, issue 6, 2015.

[13] Bjerva, Johannes, and Robert Östling, "ResSim at SemEval-2017 Task 1: Multilingual word representations for semantic textual similarity", *In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 154-158, 2017.

[14] Elekes, Ábel, Martin Schäler, and Klemens Böhm, "On the Various Semantics of Similarity in Word Embedding Models", KIT – In Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries, IEEE Press, pp. 139-148, 2017

[15] Huda, Kohei Hayashi, and Danushka Bollegala, "An Optimality Proof for the PairDiff operator for Representing Relations between Words", *arXiv preprint arXiv*:1709.06673, 2017.

[16] Ranjbar, Niloofar, Fatemeh Mashhadirajab, and Mehrnoush Shamsfard, "Mahtab at SemEval-2017 Task 2: Combination of Corpus-based and Knowledge-based Methods to Measure Semantic Word Similarity", *In Proceedings of the 11th International Workshop on Semantic Evaluation*, (SemEval-2017), pp. 256-260, 2017.

[17] Recski, Gábor, Eszter Iklódi, Katalin Pajkossy, and Andras Kornai, A., "Measuring semantic similarity of words using concept networks",

In Proceedings of the 1st Workshop on Representation Learning for NLP, pp. 193-200. 2016.

[18] Alvarez, Marco A., and SeungJin Lim. "A graph modeling of semantic similarity between words." In *null*, pp. 355-362. IEEE, 2007.

[19] Aditya Parameswaran, Anand Rajaraman, Hector Garcia-Molina, "Towards The Web Of Concepts: Extracting Concepts from Large Datasets", *ACM*, VLDB '10, Singapore, September 13-17, 2010

[20] Boucenna, Fateh, Omar Nouali, and Samir Kechid, "Concept-based Semantic Search over Encrypted Cloud Data", *In Proceedings of the 12th International Conference on Web Information Systems and Technologies*. Rome, Italy, pp. 235-242, 2016.

[21] Egozi, Ofer, Shaul Markovitch, and Evgeniy Gabrilovich, "Concept-Based Information Retrieval Using Explicit Semantic Analysis", *ACM Transactions on Information Systems.* Vol 29 Issue 2, Article No. 8. New York, NY, USA. pp. 1–34, April 2011.

[22] Fonseca, B.M., Golgher, P.B., Pôssas, B., Ribeiro-Neto, B.A., and Ziviani, N., "Concept-based query expansion", *In: CIKM*, 696–703, 2005.

[23] Lu, Yi-Jie, Phuong Anh Nguyen, Hao Zhang, and Chong-Wah Ngo, "Concept-Based Interactive Search System", *International Conference on Multimedia Modeling MMM*, Springer, Cham, pp. 463-468, 2017.

[24] Metzler, Donald, and W. Bruce Croft. "Latent concept expansion using markov random fields." *In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 311-318. ACM, 2007.

[25] Sendhil kumar, S., and T. V. Geetha., "Concept based Personalized Web Search", *TMRF, Advances in Semantic Computing*, Vol. 2. pp. 79- 102, 2010.

[26] Song, Yangqiu, Haixun Wang, Zhongyuan Wang, Hongsong Li, and Weizhu Chen, "Short text conceptualization using a probabilistic knowledgebase", *In: IJCAI,* , Vol 3, pp. 2330-2336, ISBN: 978-1-57735-515-1, 2011.

[27] Wang, Yue, Hongsong Li, Haixun Wang, and Kenny Q. Zhu, "Concept-Based Web Search", *International Conference on Conceptual Modeling*, Springer, Berlin, pp. 449-462, 2012.

[28] Wang, Zhongyuan, Kejun Zhao, Haixun Wang, Xiaofeng Meng, and Ji-Rong Wen, ”Query Understanding through Knowledge-Based Conceptualization”, Proceeding *IJCAI Microsoft Research*, pp. 3264-3270, July 2015.

[29] Wang, Fang, Zhongyuan Wang, Zhoujun Li, and Ji-Rong Wen. "Concept-based short text classification and ranking." *In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pp. 1069-1078. ACM, 2014.

[30] Zhang, Lanbo, “Interactive Retrieval Based on Wikipedia Concepts”, *Arxiv Preprint arXiv*. 1412.8281, 2014.

[31] Bjerva, Johannes, and Robert Östling. "ResSim at SemEval-2017 Task 1: Multilingual word representations for semantic textual similarity." *In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 154-158. 2017.

[32] Recski, Gábor, Eszter Iklódi, Katalin Pajkossy, and Andras Kornai. "Measuring semantic similarity of words using concept networks." *In Proceedings of the 1st Workshop on Representation Learning for NLP*, pp. 193-200. 2016.

[33] https://www.microsoft.com/en-us/research/project/probase/

[34] https://en.wikipedia.org/wiki/Archie_search_engine

[35] https://en.wikipedia.org/wiki/Veronica &Jughead

[36] http://scg.unibe.ch/archive/software/w3catalog/W3CatalogHistory.html

[37] https://en.wikipedia.org/wiki/W3Catalog

[38] https://en.wikipedia.org/wiki/JumpStation

[39] https://en.wikipedia.org/wiki/World_Wide_Web_Wanderer

[40] http://thesearchenginearchive.wikia.com/wiki/Aliweb

[41] http://www.sciencedaily.com/terms/web_crawler.htm

[42] https://en.wikipedia.org/wiki/MetaCrawler

[43] http://malwaretips.com/blogs/remove-mywebsearch/

[44] http://www.livinginternet.com/w/wu_sites_lycos.htm

[45] http://searchenginewatch.com/sew/news/2047873/inktomi-debuts-self-serve-paid-inclusion

[46] https://en.wikipedia.org/wiki/Infoseek

[47] https://en.wikipedia.org/wiki/Excite

[48] http://searchenginewatch.com/sew/study/2067828/altavistas-search-by-language-feature
[49] http://www.searchengineshowdown.com/features/yahoo/review.html
[50] https://en.wikipedia.org/wiki/Yahoo!_Search
[51] https://en.wikipedia.org/wiki/AOL
[52] http://www.msn.com/en-in/
[53] https://en.wikipedia.org/wiki/Dogpile
[54] http://investor.blucora.com/releasedetail.cfm?ReleaseID=166325
[55] http://chj.tbe.taleo.net/chj04/ats/careers/requisition.jsp?org=INFOSPACE & cws=1& rid=181
[56] https://en.wikipedia.org/wiki/HotBot
[57] http://www.searchengineshowdown.com/features/hotbot/review.html
[58] https://en.wikipedia.org/wiki/Wow!_(online_service)
[59] https://en.wikipedia.org/wiki/Ask.com
[60] http://www.searchengineshowdown.com/features/ask/review.html
[61] https://en.wikipedia.org/wiki/Daum_(web_portal)
[62] http://www.search-marketing.info/search-engines/price-per-click/overture.htm
[63] https://en.wikipedia.org/wiki/Yandex_Search
[64] https://en.wikipedia.org/wiki/Google_Search#calculator
[65] http://www.telegraph.co.uk/technology/google/10346736/Google-search-15-hidden-features.html
[66] https://en.wikipedia.org/wiki/AlltheWeb
[67] http://www.seochat.com/c/a/marketing/web-directories/teoma-the-superior-search-engine/
[68] https://en.wikipedia.org/wiki/Baidu
[69] https://en.wikipedia.org/wiki/Live_search
[70] https://en.wikipedia.org/wiki/DuckDuckGo
[71] https://en.wikipedia.org/wiki/Aardvark_(search_engine)
[72] http://www.windowscentral.com/top-bing-features
[73] http://www.telegraph.co.uk/technology/google/6009176/Google-reveals-caffeine-a-new-faster-search engine.html
[74] http://searchengineland.com/google-instant-complete-users-guide-50136
[75] https://en.wikipedia.org/wiki/Blekko

[76] https://www.aihitdata.com/company/00D2051A/CONTENKO/history

[77] https://en.wikipedia.org/wiki/Althea

[78] http://www.searchenginejournal.com/seo-guide/google-penguin-panda-hummingbird/

[79] Tuukka Ruotsalo, Kumaripaba Athukorala, Dorota Głowacka, Ksenia Konyushkova, Antti Oulasvirta, Samuli Kaipiainen, Samuel Kaski, Giulio Jacucci, "Supporting Exploratory Search Tasks with Interactive User Modeling", *Helsinki Institute for Information Technology HIIT,* University of Helsinki, *ASIST 2013,* November 1-6, 2013

[80] https://searchenginewatch.com/tag/clusty/

[81] https://in.linkedin.com/company/lexxe-search

[82] Claudio Carpineto, Giovanni Romano, "A Survey of Automatic Query Expansion in Information Retrieval", ACM Comput.Surv. Vol 44, issue 1, Article 1, pp. 50, January 2012.

[83] Baeza-Yates & Ribeiro-Neto, "Modern Information Retrieval", ACM Press, New York, July 1999

[84] Bruno M. Fonseca, Paulo B. Golgher, Edleno S. de Moura and Nivio Ziviani, "Using Association Rules to Discover Search Engines Related Queries", LA-WEB '03 Proceeding of the first conference on Latin American Web Congress, pp. 66, ISBN: 0-7695-2058-8, November, 2003.

[85] Paolo Boldi, Francesco Bonchi and Carlos Castillo, "The Query-flow Graph: Model and Applications"CIKM'08, October 26–30, 2008, Napa Valley, California, USA

[86] Doug Beeferman and Adam Berger, "Agglomerative clustering of a search engine query log", KDD Proceeding of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 407-416, August 2000.

[87] Boubacar, Abdoulahi, and Zhendong Niu. "Concept Based Query Expansion." In Semantics, Knowledge and Grids (SKG), *Ninth International Conference,* pp. 198-201. IEEE, 2013.

[88] Pilehvar, Mohammad Taher, David Jurgens, and Roberto Navigli. "Align, disambiguate and walk: A unified approach for measuring semantic similarity." *In Proceedings of the 51st Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), vol. 1, pp. 1341-1351. 2013.

[89] Severyn, Aliaksei, Massimo Nicosia, and Alessandro Moschitti. "Learning semantic textual similarity with structural representations." *In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers), vol. 2, pp. 714-718. 2013.

[90] Specia, Lucia, Sujay Kumar Jauhar, and Rada Mihalcea. "Semeval-2012 task 1: English lexical simplification." In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, Association for Computational Linguistics, pp. 347-355, 2012.

[91] Jurgens, David A., Peter D. Turney, Saif M. Mohammad, and Keith J. Holyoak. "Semeval-2012 task 2: Measuring degrees of relational similarity." In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, Association for Computational Linguistics, pp. 356-364, 2012.

[92] Cilibrasi, Rudi L., and Paul MB Vitanyi, "The google similarity distance", *IEEE Transactions on knowledge and data engineering*, Vol 19, Issue 3, Pages 370-383, 2007.

[93] Chen, Hsin-Hsi, Ming-Shun Lin, and Yu-Chuan Wei, "Novel association measures using web search with double checking", *In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics* , Association for Computational Linguistics, pp. 1009-1016, July, 2007.

[94] Rada, Roy, Hafedh Mili, Ellen Bicknell, and Maria Blettner, "Development and application of a metric on semantic nets", *IEEE Transactions on systems*, man, and cybernetics, Vol 19, issue 1, pp. 17-30, 1989

[95] Resnik, Philip, "Using Information Content to Evaluate Semantic Similarity in a Taxonomy", *Proc. 14th Int'l Joint Conf. Aritificial Intelligence*. Sun Microsystems Laboratories Two Elizabeth Drive Chelmsford, MA 01824-4195 USA, 1995.

[96] Brown Corpus(1991) [documents] Retrieved from http://www.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/corpora/list/private/brown/brown.html

[97] WordNet (2005), [Lexical Resource] Retrieved from http://wordnetweb.princeton.edu/perl/webwn

[98] Bollacker, Kurt, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. "Freebase: a collaboratively created graph database for structuring human knowledge." In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, ACM, pp. 1247-1250, 2008.

[99] http//www.wikipedia.com

[100] Strube, Michael, and Simone Paolo Ponzetto. "WikiRelate! Computing semantic relatedness using Wikipedia." In *AAAI*, vol. 6, pp. 1419-1424. 2006.

[101] Hu, Jian, Lujun Fang, Yang Cao, Hua-Jun Zeng, Hua Li, Qiang Yang, and Zheng Chen. "Enhancing text clustering by leveraging Wikipedia semantics." In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 179-186. ACM, 2008.

[102] Chaudhuri, Surajit, Venkatesh Ganti, and Dong Xin. "Exploiting web search to generate synonyms for entities." In *Proceedings of the 18th international conference on World wide web*, pp. 151-160. ACM, 2009.

[103] Malekian, Azarakhsh, Chi-Chao Chang, Ravi Kumar, and Grant Wang. "Optimizing query rewrites for keyword-based advertising." In *Proceedings of the 9th ACM conference on Electronic commerce*, pp. 10-19. ACM, 2008.

[104] Dey, Debabrata, Sumit Sarkar, and Prabuddha De. "A distance-based approach to entity reconciliation in heterogeneous databases." *IEEE Transactions on Knowledge & Data Engineering*, pp. 567-582, 2002.

[105] Dong, Xin, Alon Halevy, and Jayant Madhavan. "Reference reconciliation in complex information spaces." In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pp. 85-96. ACM, 2005.

[106] Benjelloun, Omar, Hector Garcia-Molina, David Menestrina, Qi Su, Steven Euijong Whang, and Jennifer Widom. "Swoosh: a generic approach to entity resolution." *The VLDB Journal—The International Journal on Very Large Data Bases*, vol. 18, no. 1, pp. 255-276, 2009.

[107] Bhattacharya, Inderjit and Lise Getoor, “Collective Entity Resolution in Relational Data”, *ACM Trans. Knowledge Discovery from Data (TKDD),* pp. 1 1–36, 2007.

[108] Kaushik Chakrabarti, Surajit Chaudhuri, Tao Cheng and Dong Xin, “A Framework for Robust Discovery of Entity Synonyms”, *In KDD '12 Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, Beijing, China August 2012.

[109] Surajit Chaudhuri, Venkatesh Ganti and Dong Xin, ”Mining Document Collections to Facilitate Accurate Approximate Entity Matching”, *In Pprceeding VLDB,* August 2009.

[110] Srikantiah K. C., Roopa M. S., Krishna Kumar N., Tejaswi V., Venugopal K. R. and Patnaik L. M, “Automatic Discovery of Synonyms from the Web based on Inbound Anchor Text”, *ICDMW* , © Elsevier Publications. pp. 130–139. 2013.

[111] Xiang Ren and Tao Cheng, ”Synonym Discovery for Structured Entities on Heterogeneous Graphs”, *In WWW '15 Companion Proceedings of the 24th International Conference on World Wide Web*, Florence, Italy, May 18 - 22, 2015

[112] Ponzetto, Simone Paolo. and Michael Strube, “Deriving a large-scale taxonomy from wikipedia”, *In Proceeding of Association for the Advancement of Artificial Intelligence AAAI*, vol. 7, pp. 1440-1445, 2007.

[113] Lenat, Douglas B. and Ramanathan V. Guha, “Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project”, Building large knowlwdge based system, *Addison-Wesley*, ISBN:0201517523, 1989.

[114] Suchanek, Fabian M., Gjergji Kasneci, and Gerhard Weikum. "Yago: a core of semantic knowledge." *In Proceedings of the 16th international conference on World Wide Web*, pp. 697-706. ACM, 2007.

[115] Etzioni, Oren, Michael Cafarella, Doug Downey, Stanley Kok, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Deniel S. Weld and Alexander Yates, A., “Web-scale information extraction in knowitall”, *In Proceeding of 3rd international conference on World Wide Web, WWW,* pp. 100–110, ACM, 2004.

[116] Banko, Michele, Michael J. Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. "Open information extraction from the web." In *IJCAI*, vol. 7, pp. 2670-2676, 2007.

[117] Singh, Push, Thomas Lin, Erik T. Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. "Open Mind Common Sense: Knowledge acquisition from the general public." In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pp. 1223-1237. Springer, Berlin, Heidelberg, 2002.

[118] Carlson, Andrew, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr, and Tom M. Mitchell. "Toward an architecture for never-ending language learning", *In AAAI*, vol. 5, pp.3, 2010.

[119] Auer, Sören, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. "Dbpedia: A nucleus for a web of open data." In *The semantic web*, pp. 722-735. Springer, Berlin, Heidelberg, 2007.

[120] Wang, Zhongyuan, Jiuming Huang, Hongsong Li, Bin Liu, Bin Shao, Haixun Wang, Jingjing Wang et al. "Probase: a Universal Knowledge Base for Semantic Search", *Microsoft Research Area*, Pages 1-3, May 2011.

[121] Wu, Wentao, Hongsong Li, Haixun Wang, and Kenny Q. Zhu, "Probase : a probabilistic taxonomy for text understanding", *In Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, ISBN: 978-1-4503-1247, pp. 481-492, ACM, May 2012.

[122] Song, Yangqiu, Haixun Wang, Zhongyuan Wang, Hongsong Li, and Weizhu Chen. "Short text conceptualization using a probabilistic knowledgebase." In *Proceedings of the twenty-second international joint conference on artificial intelligence-volume volume three*, pp. 2330-2336. AAAI Press, 2011.

[123] Lee, Taesung, Zhongyuan Wang, Haixun Wang, and Seung-won Hwang. "Web scale taxonomy cleansing." *Proceedings of the VLDB Endowment* 4, no. 12, pp. 1295-1306, 2011.

[124] Lu, Y.J., Nguyen, P.A., Zhang, H., Ngo, C.W., "Concept-Based Interactive Search System", *International Conference on Multimedia Modeling MMM*, Pages 463-468, 2017.

[125] Liu, Xueqing, Yangqiu Song, Shixia Liu, and Haixun Wang. "Automatic taxonomy construction from keywords." In *Proceedings of the 18th ACM*

SIGKDD international conference on Knowledge discovery and data mining, pp. 1433-1441. ACM, 2012.

[126] Egozi, Ofer, Shaul Markovitch, and Evgeniy Gabrilovich., "Concept-Based Information Retrieval Using Explicit Semantic Analysis", *ACM Transactions on Information Systems*. Vol 29 Issue 2, Article No. 8. New York, NY, USA. pp. 1–34, April 2011.

[127] Sendhilkumar, S., and T. V. Geetha, "Concept based Personalized Web Search", *Advances in Semantic Computing*, Anna University, Chennai, Vol. 2. Pages 79- 102, 2010.

[128] Fonseca, Bruno M., Paulo Golgher, Bruno Pôssas, Berthier Ribeiro-Neto, and Nivio Ziviani. "Concept-based interactive query expansion." In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pp. 696-703. ACM, 2005.

[129] Naphade, Milind R., and John R. Smith. "On the detection of semantic concepts at TRECVID." In *Proceedings of the 12th annual ACM international conference on Multimedia*, pp. 660-667. ACM, 2004.

[130] Metzler, Donald, and W. Bruce Croft. "Latent concept expansion using markov random fields." In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 311-318. ACM, 2007.

[131] Boucenna, Fateh, Omar Nouali, and Samir Kechid, "Concept-based Semantic Search over Encrypted Cloud Data", *In Proceedings of the 12th International Conference on Web Information Systems and Technologies*. Rome, Italy, Pages 235-242, 2016.

[132] https://www.pearson.com/.../Klir-Fuzzy...Fuzzy-Logic-Theory-and-Applications/PGM

[133] Bollegala, Danushka, Yutaka Matsuo, and Mitsuru Ishizuka, "Measuring semantic similarity between words using web search engines" ,www 7 Pages 757-766, 2007.

[134] Elekes, Ábel, Martin Schäler, and Klemens Böhm. "On the various semantics of similarity in word embedding models." In *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries*, pp. 139-148. IEEE Press, 2017.

[135] Coca Corpus(1990), [Document set] Retrieved from http://corpus.byu.edu/coca/

[136] BNC Corpus (1980), [Document set] Retrieved from http://corpus.byu.edu/bnc/

[137] Wikipedia Corpus (2006), [Document set] Retrieved from http://corpus.byu.edu/wiki/

[138] GloWbE Corpus (2005), [Document set] Retrieved from http://corpus.byu.edu/glowbe/

[139] Semantic Similarity Toolkit(2004), [Similarity ToolKit] Retrieved from http://swoogle.umbc.edu/SimService/

[140] Stanford Webbase corpus (2007) Retrieved from http://ebiquity.umbc.edu/resource/html/id/351

[141] LDC English Gigaword corpus (2003) Retrieved from https://catalog.ldc.upenn.edu/ldc2003t05

[142] Thada, Vikas, and Vivek Jaglan. "Comparison of jaccard, dice, cosine similarity coefficient to find best fitness value for web retrieved documents using genetic algorithm." *International Journal of Innovations in Engineering and Technology* 2, no. 4, pp.202-205, 2013.

[143] https://archive.org/details/AOL_search_data_leak_2006

[144] Natthawut Kertkeidkachorn, Ryutaro Ichise, “T2KG: An End-to-End System for Creating Knowledge Graph from Unstructured Text”, *The AAAI-17 Workshop on Knowledge-Based Techniques for Problem Solving and Reasoning* WS-17-12, 2007

[145] https://concept.research.microsoft.com/Home/Introduction

[146] Priyadarshini, R., and Latha Tamilselvan. "Document clustering based on keyword frequency and concept matching technique in Hadoop." *International Journal of Scientific & Engineering Research* 5, no. 5, 2014.

[147] Zhang, Jiuling, Beixing Deng, and Xing Li. "Concept based query expansion using wordnet." In Proceedings of the 2009 international e-conference on advanced science and technology, pp. 52-55. IEEE Computer Society, 2009.

APPENDIX-A

The methodology proposed is Chapter III is corpus centric (starting from the corpus analysis to build the Next Generation Lexical Resource using contexts from corpuses). Four different corpuses are used to extract the set of contexts for both the input query word and its synonyms obtained from the lexical resource. These corpuses can handle complex queries typically in two or three seconds. Finally, the relational database design of these corpuses allows a range of queries that we believe is unmatched by any other architecture for large corpora. In order to search the context, input word (adjective) is used along with an asterisk (i.e. beautiful*) on the interface of the online available corpus. The snapshot is shown below in Fig. A.1.

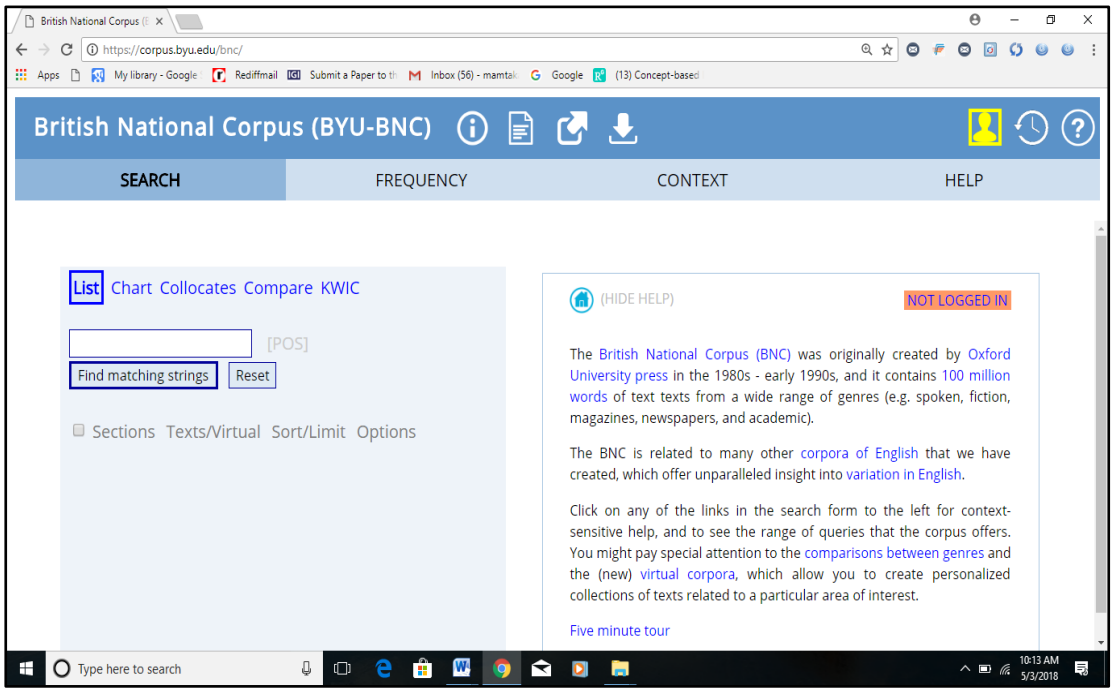


Fig. A.1 Snapshot of the British National Corpus (BNC)

In Fig. A.2 area A depicts the various contexts of word “beautiful” extracted from corpus BNC and area B depicts its frequency in the corpus.

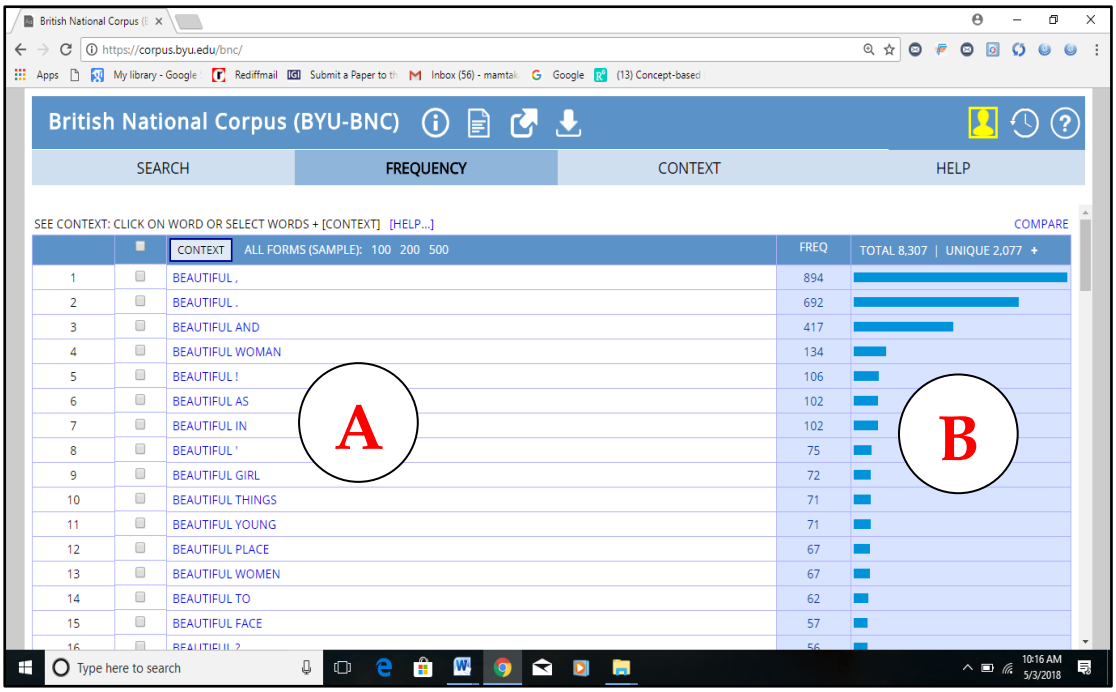


Fig. A.2 Snapshot of the context associated with the word beautiful in British National Corpus (BNC)

Fig. A.3 depicts the only large, balanced corpus of contemporary American English having 520 million words related to different domains including newspaper and magazine.

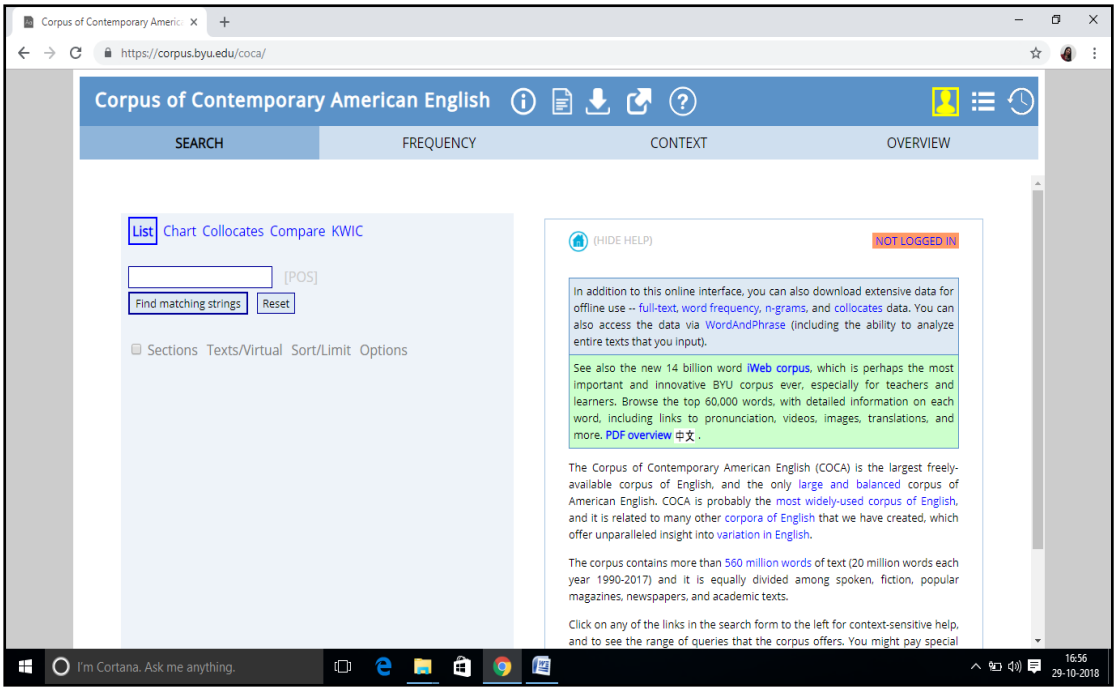


Fig. A.3 Snapshot of the Corpus of Contemporary American English (COCA)

Fig. A.4 shows the contexts for the word beautiful along with frequency of its usage in the corpus.

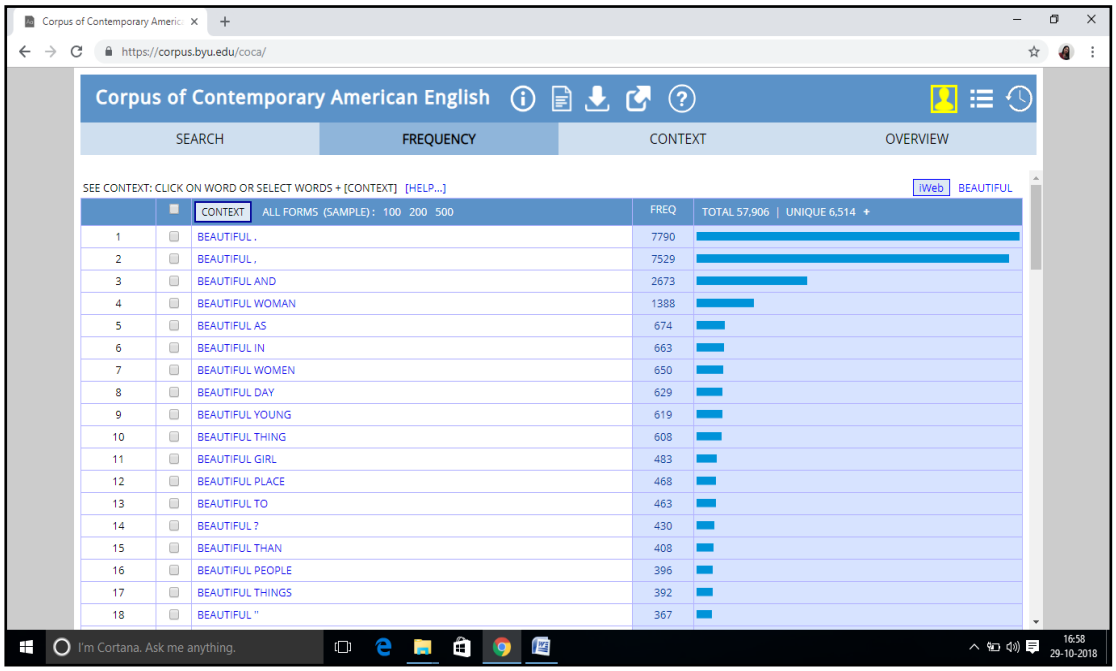


Fig. A.4 Snapshot of the Corpus of Contemporary American English (COCA) shows the context for the word beautiful

Fig. A.5 shows the interface of a huge Wikipedia corpus having 1.9 billion words and 4.4 million documents related to microbiology, economics, basketball, Buddhism, or thousands of other topics.

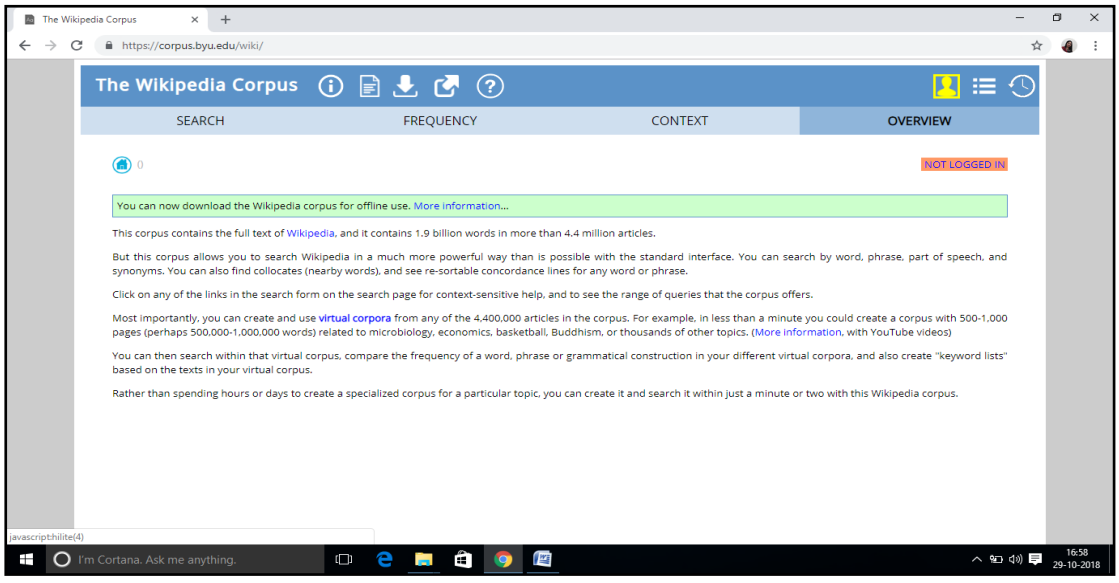


Fig. A.5 Snapshot of the Wikipedia Corpus

Fig. A.6 depicts the interface of the Wikipedia corpus.

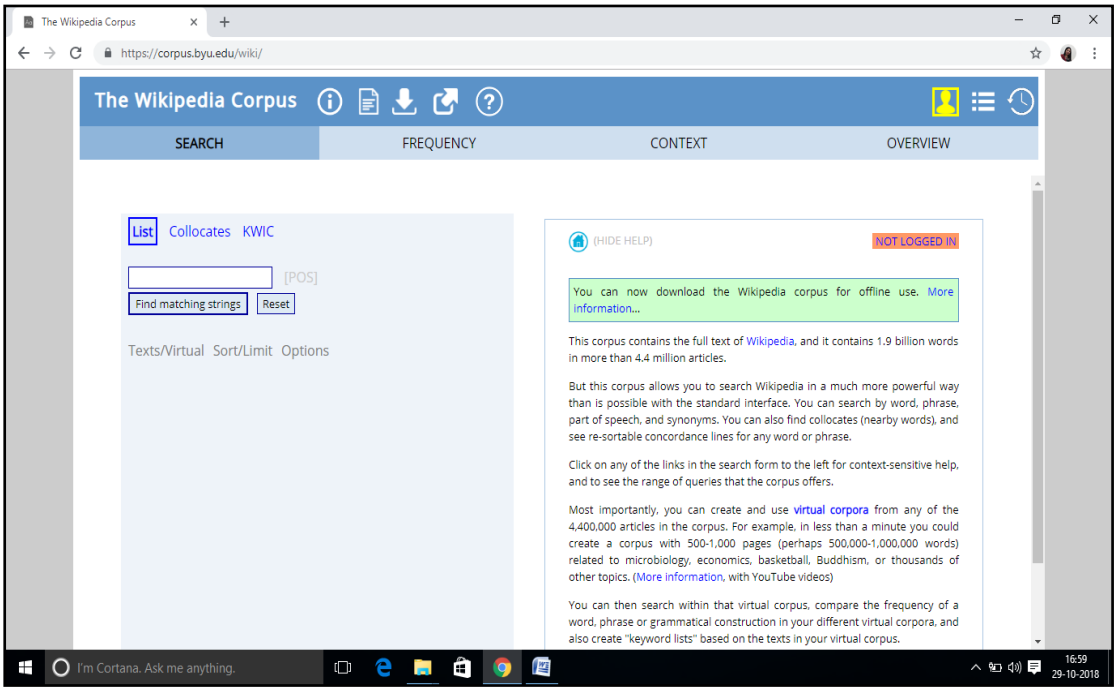


Fig. A.6 Snapshot of the Wikipedia Corpus

Interface for the other huge GLOWBE corpus is shown in Fig. A.7 which contain any type of data related to newspaper, magazines and academic.

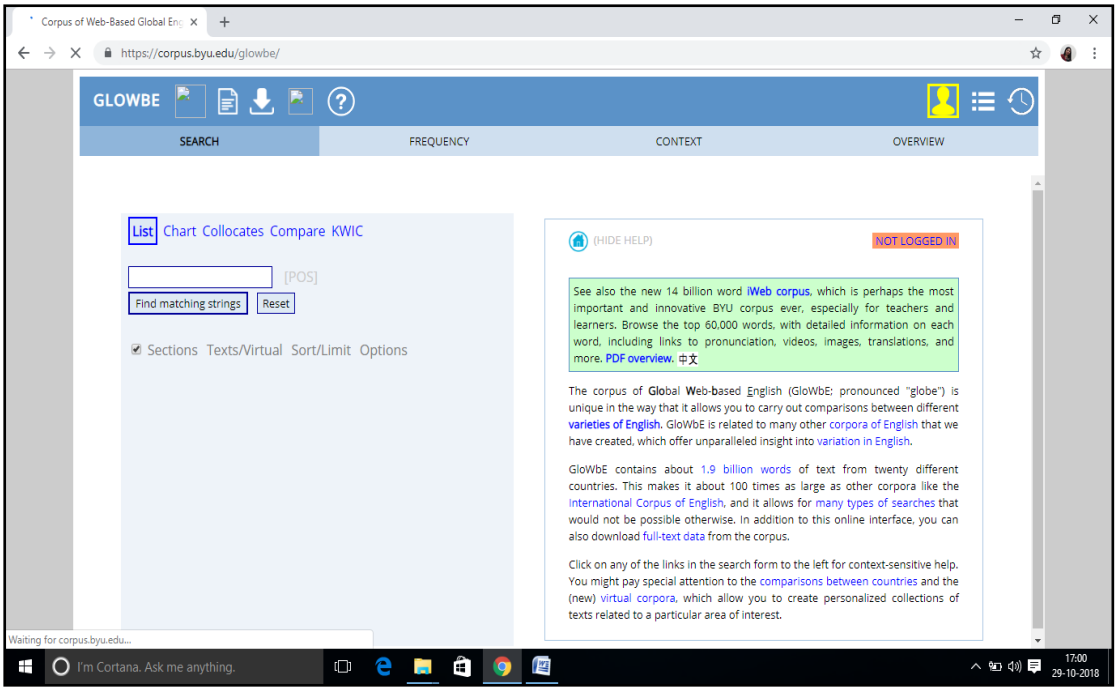
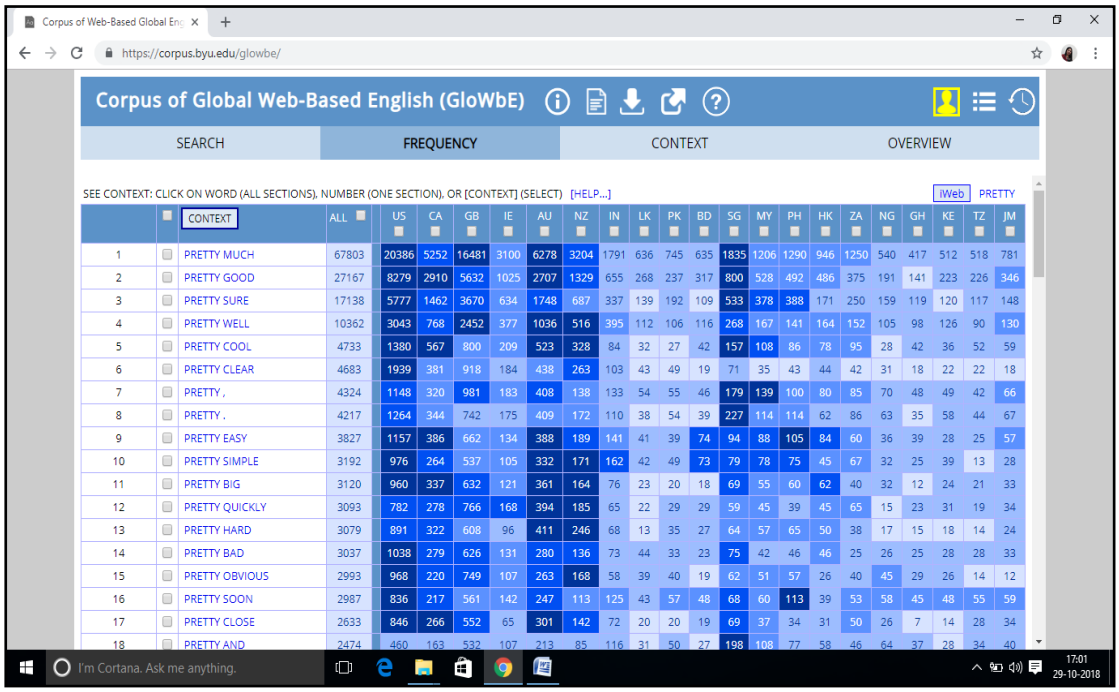


Fig. A.7 Snapshot of the GLOWBE Corpus

Fig. A.8 depicts the snapshot of the GLOWBE Corpus for the word Pretty along with associated frequency in different countries.



	CONTEXT	ALL	US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
1	PRETTY MUCH	67803	20386	5252	16481	3100	6278	3204	1791	636	745	635	1835	1206	1290	946	1250	540	417	512	518	781
2	PRETTY GOOD	27167	8279	2910	5632	1025	2707	1329	655	268	237	317	800	528	492	486	375	191	141	223	226	346
3	PRETTY SURE	17138	5777	1462	3670	634	1748	687	337	139	192	109	533	378	388	171	250	159	119	120	117	148
4	PRETTY WELL	10362	3043	768	2452	377	1036	516	395	112	106	116	268	167	141	164	152	105	98	126	90	130
5	PRETTY COOL	4733	1380	567	800	209	523	328	84	32	27	42	157	108	86	78	95	28	42	36	52	59
6	PRETTY CLEAR	4683	1939	381	918	184	438	263	103	43	49	19	71	35	43	44	42	31	18	22	22	18
7	PRETTY ,	4324	1148	320	981	183	408	138	133	54	55	46	179	139	100	80	85	70	48	49	42	66
8	PRETTY .	4217	1264	344	742	175	409	172	110	38	54	39	227	114	114	62	86	63	35	58	44	67
9	PRETTY EASY	3827	1157	386	662	134	388	189	141	41	39	74	94	88	105	84	60	36	39	28	25	57
10	PRETTY SIMPLE	3192	976	264	537	105	332	171	162	42	49	73	79	78	75	45	67	32	25	39	13	28
11	PRETTY BIG	3120	960	337	632	121	361	164	76	23	20	18	69	55	60	62	40	32	12	24	21	33
12	PRETTY QUICKLY	3093	782	278	766	168	394	185	65	22	29	29	59	45	39	45	65	15	23	31	19	34
13	PRETTY HARD	3079	891	322	608	96	411	246	68	13	35	27	64	57	65	50	38	17	15	18	14	24
14	PRETTY BAD	3037	1038	279	626	131	280	136	73	44	33	23	75	42	46	46	25	26	25	28	28	33
15	PRETTY OBVIOUS	2993	968	220	749	107	263	168	58	39	40	19	62	51	57	26	40	45	29	26	14	12
16	PRETTY SOON	2987	836	217	561	142	247	113	125	43	57	48	68	60	113	39	53	58	45	48	55	59
17	PRETTY CLOSE	2633	846	266	552	65	301	142	72	20	20	19	69	37	34	31	50	26	7	14	28	34
18	PRETTY AND	2474	460	163	532	107	213	85	116	31	50	27	198	108	77	58	46	64	37	28	34	40

Fig. A.8 Snapshot of the GLOWBE Corpus for the word Pretty

These corpuses are publically available which anyone can use for free. The reason for choosing these four corpuses is their ability to store data belonging to multiple domains.

APPENDIX-B

Entity synonym finding technique based on query log and web data discussed in Section 4.5 of chapter IV is implemented. The proposed technique is implemented as a standalone tool. AOL search data release (20M queries, 650K users, 3 months) is used to extract the entity synonyms from static web and then dynamic web is also explored to get the rich and relevant set of entity synonyms.

Some minimum hardware, software and database requirements are essential for efficiently working of the proposed approach. The minimum system requirement includes:

(i) Hardware Requirements:

- Processor: Intel Core 2 Duo CPU T5470 @1.60Ghz
- Ram: 3 GB
- System: 32 bit/64 bit //If windows is of 32 bit then all software listed below should be of 32 bit.
- Fast internet connectivity for getting better and fast results

(ii) Software Requirements:

Software requirements include installation and setting up the environment for the following software:

- Eclipse Java Neon as it includes Maven dependency implicitly; for other eclipse version, maven set up need to be done.
- Java 8
- Jre8/Jdk 8 (setting up the environment variables of jre and jdk)
- Mysql server
- Mysql workbench
- Neo4j community edition for making graph database of entities
- Stanfordcorenlp parser and lemmatizer dependencies or jar files for pre processing task
- Various other dependencies like Mysql connector, Jsoup, common-lang3, common-codec etc are needed for running the entire project.

(iii) Database Requirements:

- Setting a local database connection and schema on Mysql workbench

- AOL query logs
- Addition of these AOL query logs to tables for processing the results

Query log is used to obtain the basic set of entity synonyms by matching the URLs returned on search interface with the URLs present in Query Log. A snapshot of the sample query log containing adequate amount of queries is shown in Fig. B.1 and B.2.

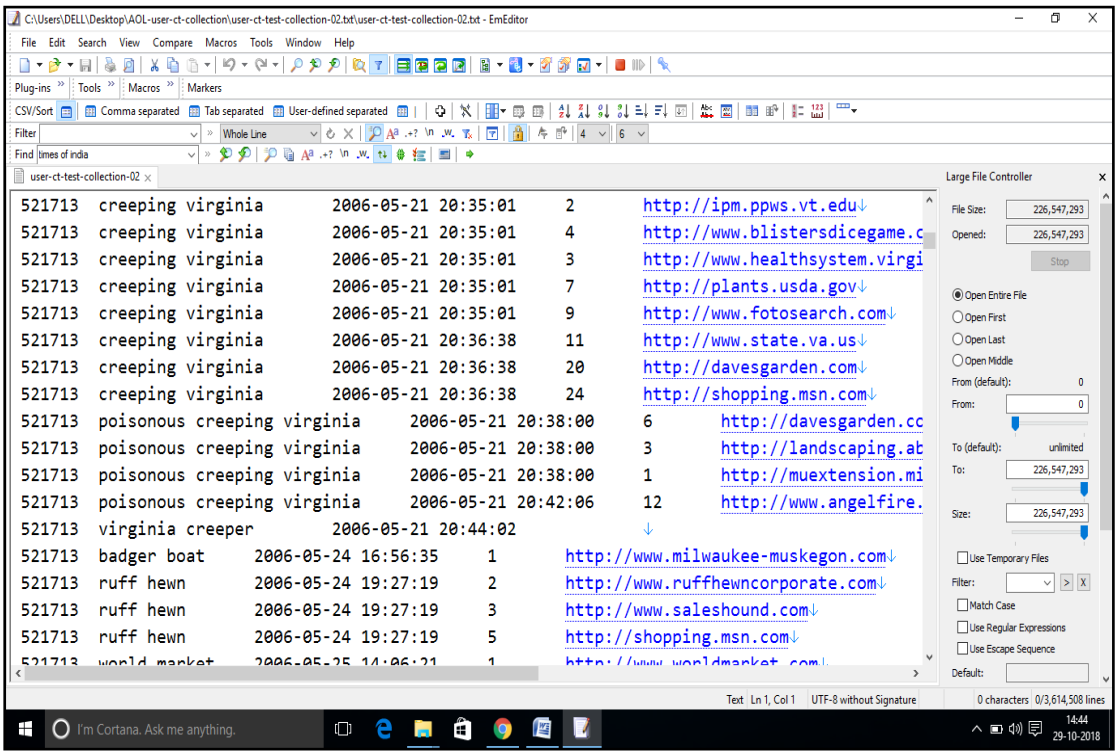


Fig. B.1 Snapshot of AOL query log

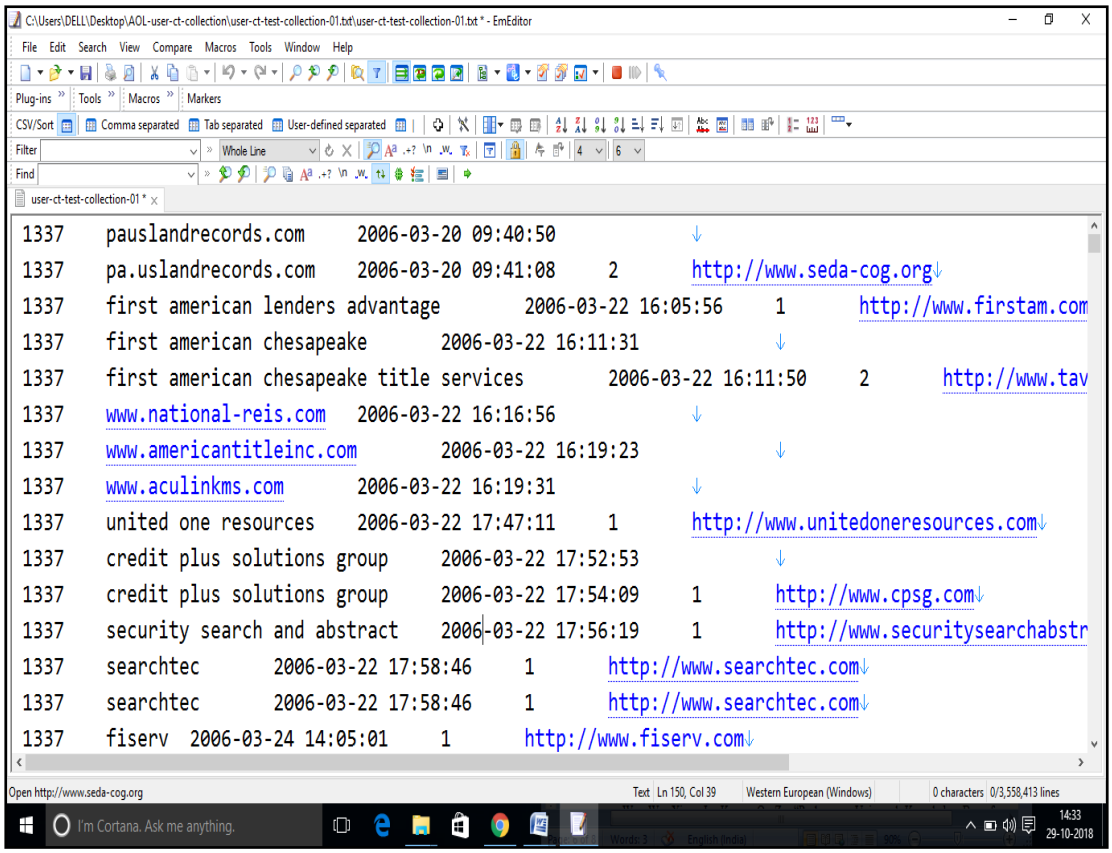


Fig. B.2 Snapshot of AOL Quey Log

The different URL clicked for the query *Times of India* is shown in Fig. B.3 & B.4.

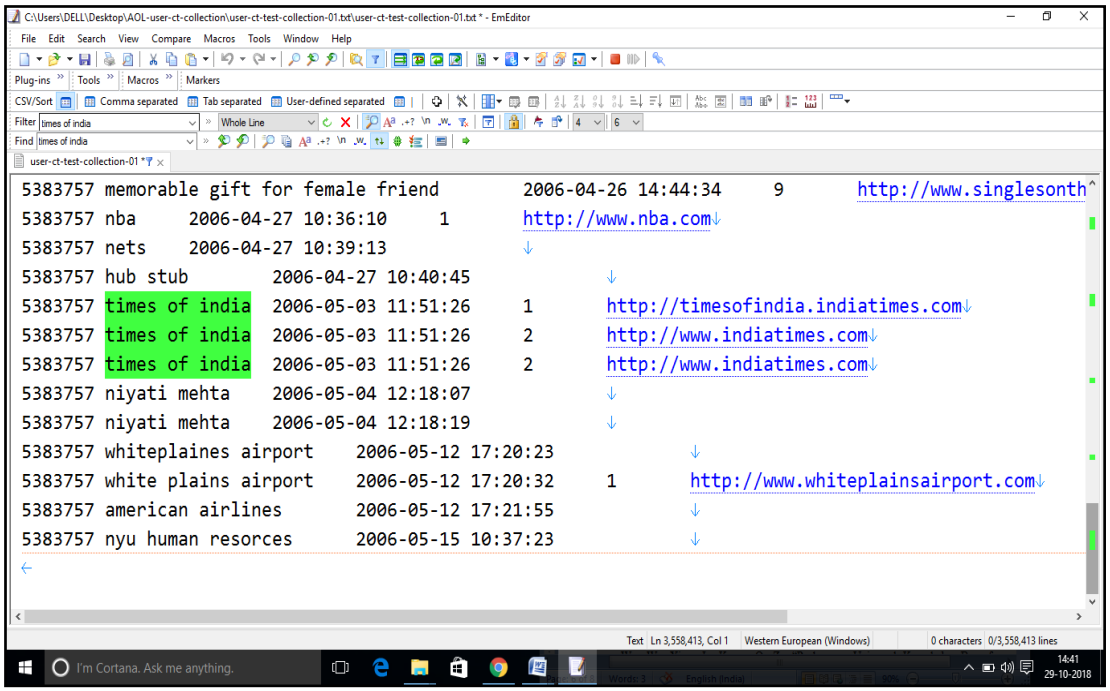


Fig. B.3 Snapshot of different URLs clicked for the query Times of India

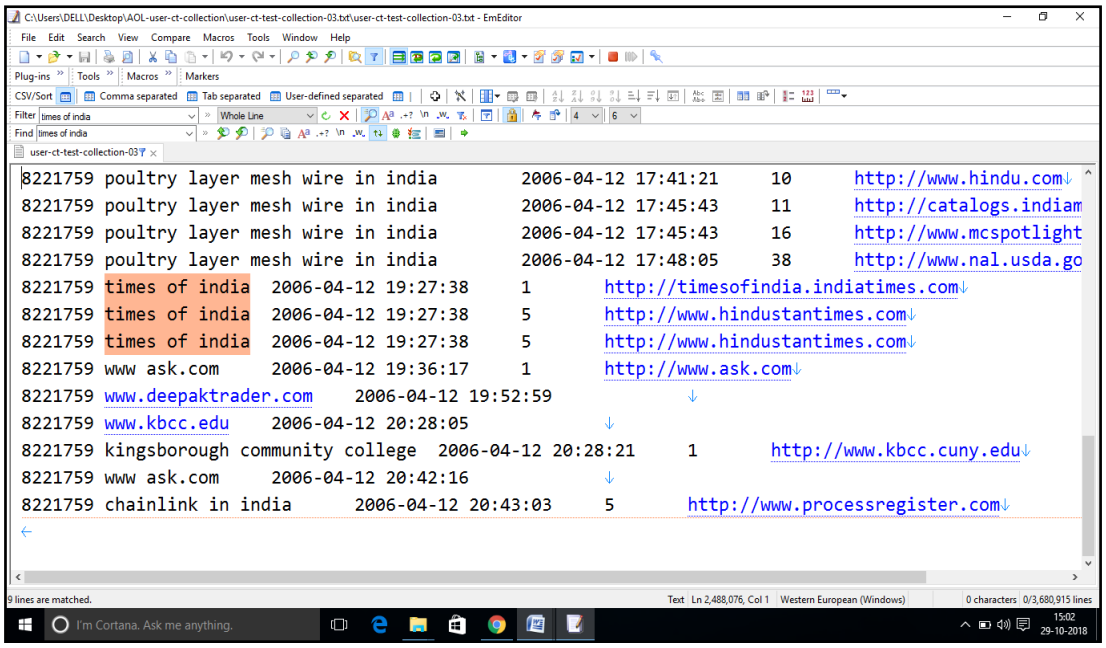


Fig. B.4 Snapshot of different URLs clicked for the query Times of India

In order to find the candidate entity synonyms from the query log, the user is first asked to enter the input query as an entity onto the search interface. The step is required to initiate the implementation of the proposed work. Then, the next step is to match the URLs returned by the search engine with the URL present in the query log to obtain the first set of candidate synonyms. Fig. B.5 shows the snapshot of the implementation work to obtain the entity synonyms from query log.

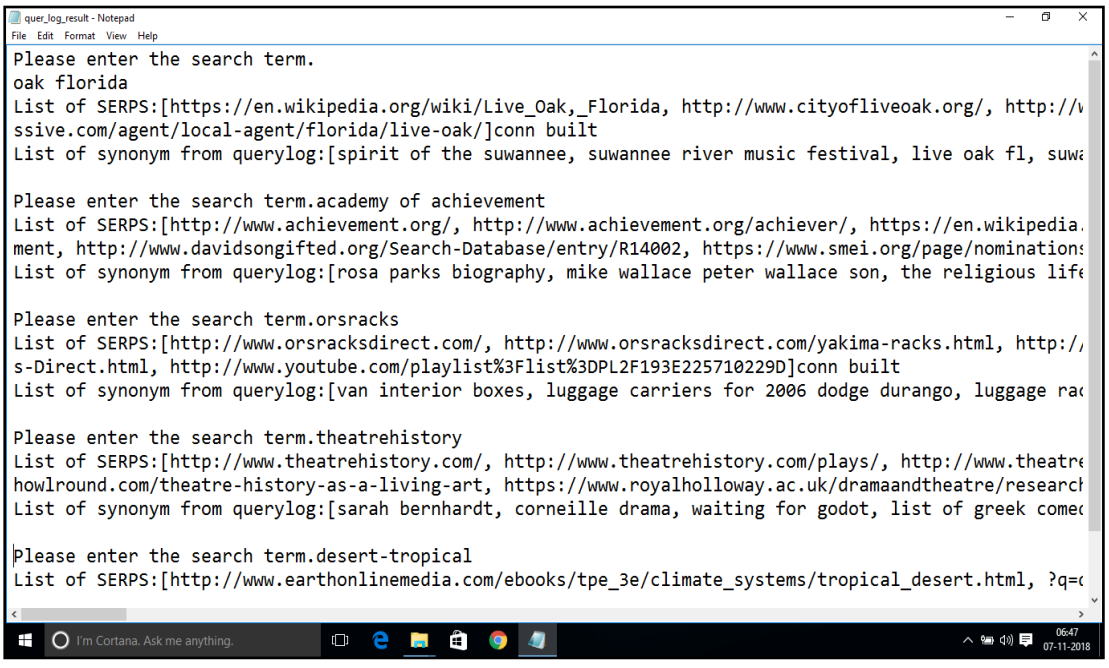


Fig. B.5 Results to find first set of entity synonyms using Query Log

The set of candidate entity synonyms are then presented back to user as shown in Fig. B.5. These basic set of candidate synonyms are then used to obtain the rich and relevant set of entity synonyms with similarity index. The snapshots are shown in Fig. B.6 & Fig. B.7. Similarity measure or similarity function is used to measure the extent of similarity between the input entity and the candidate synonyms extracted from the database. The proposed works uses WebJaccard similarity measure to accurately measure the relevance between input entity and candidate synonym, which is further used to rank the candidate synonyms.

Fig. B.6 depicts the list of entity synonyms obtained using *anchor text*.

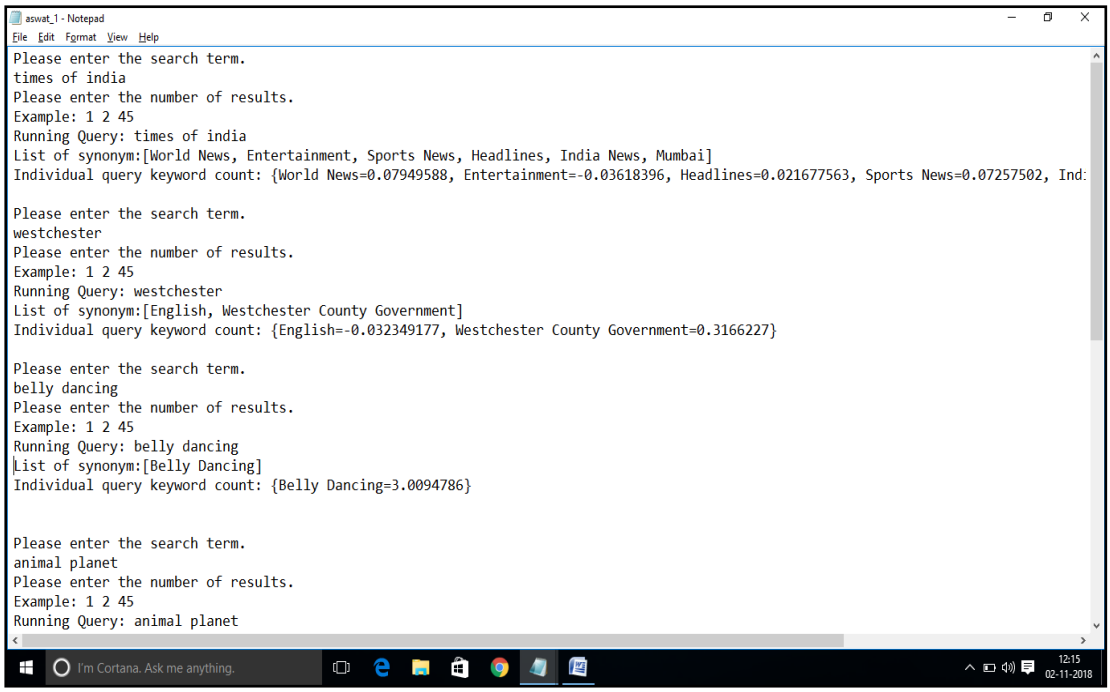


Fig. B.6 Result list of entity synonyms obtained using *Anchor Text*

Fig. B.7 & Fig. B.8 depicts the list of entity synonyms obtained using *inbound anchor text* along with the context along with the extent of similarity.

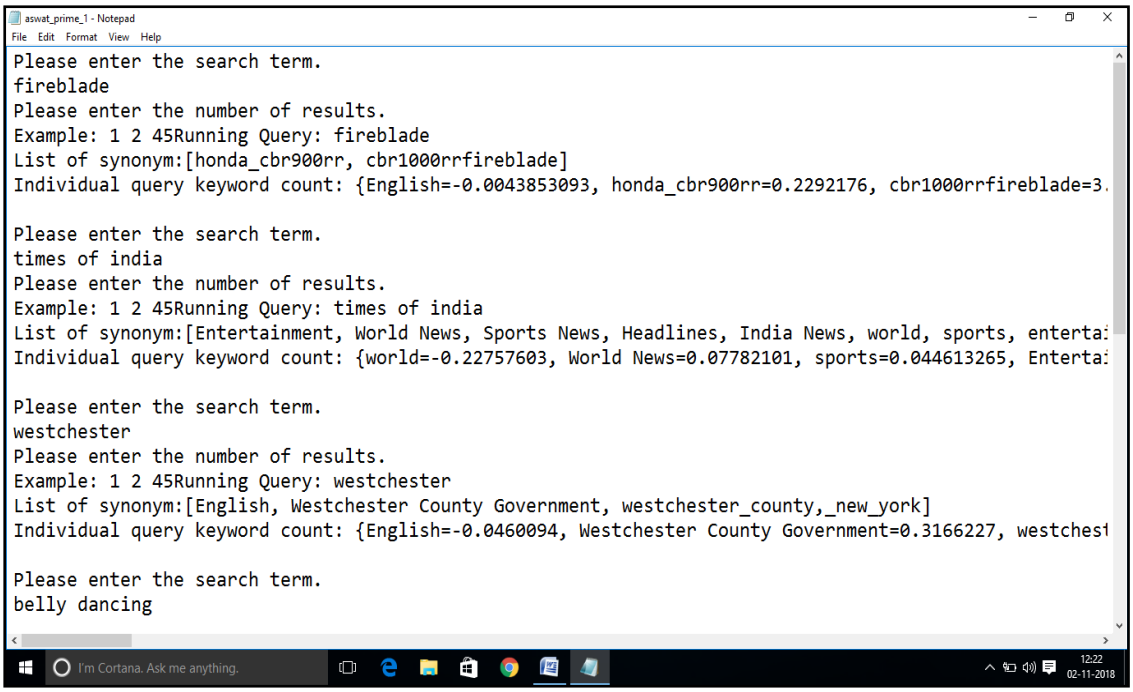


Fig. B.7 Result list of entity synonyms obtained using inbound Anchor Text + Context

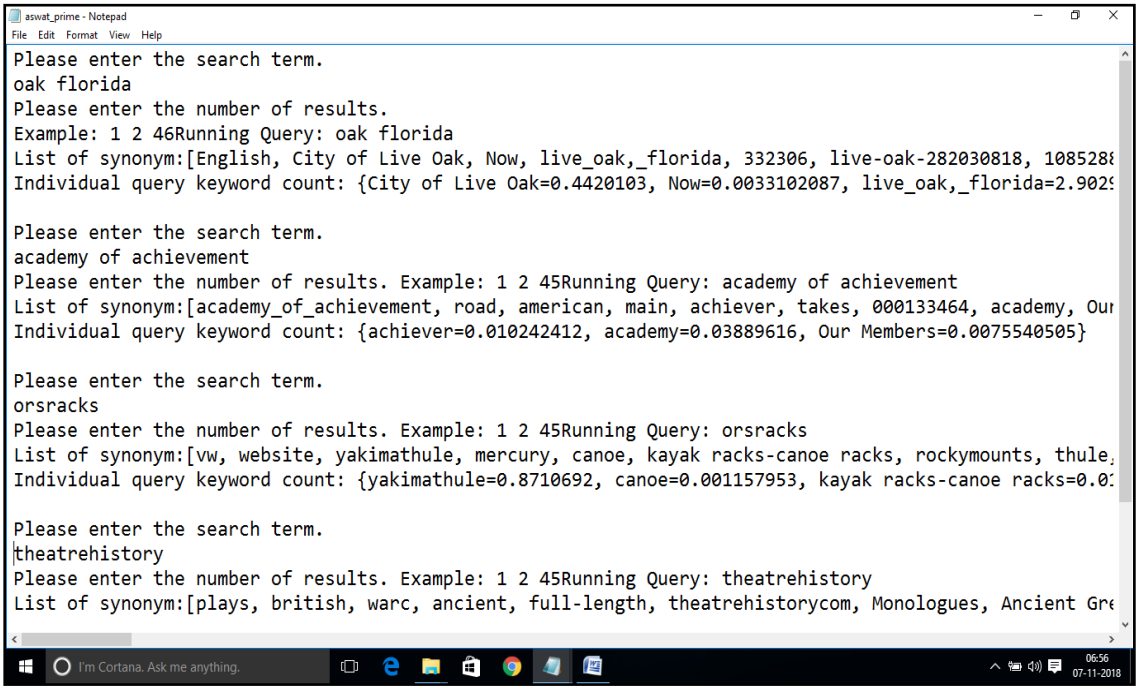


Fig. B.8 Results of the Anchor Text + Context to find entity synonyms

Fig. B.9- B.11 shows the entity synonyms obtained after implemented the proposed work. It can be noted from the implementation results of the proposed system that it is able to find more optimized set of entity synonyms both in terms of quality and quantity.

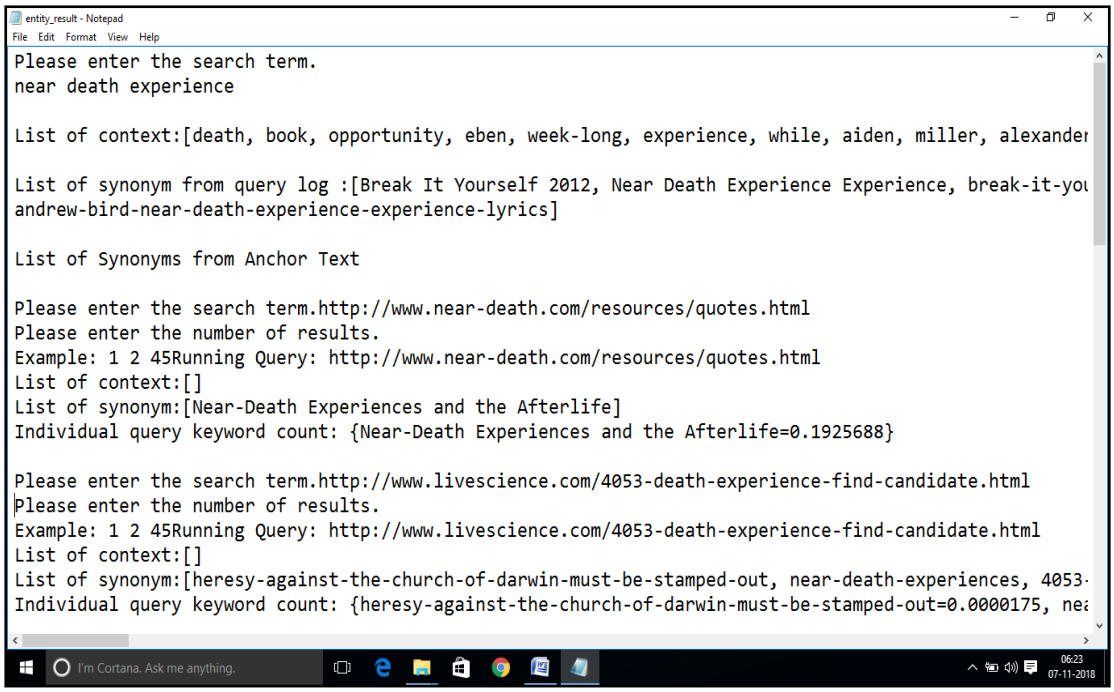


Fig. B.9 Results of the Proposed work to find entity synonyms

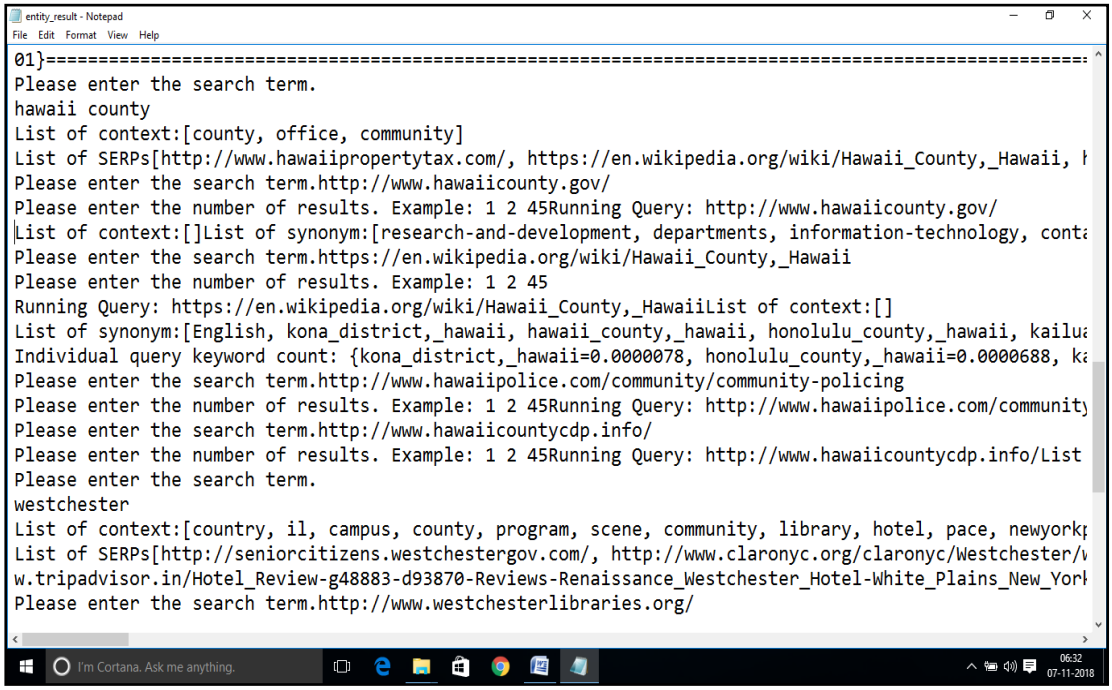


Fig. B.10 More Results of the proposed work to find entity synonyms

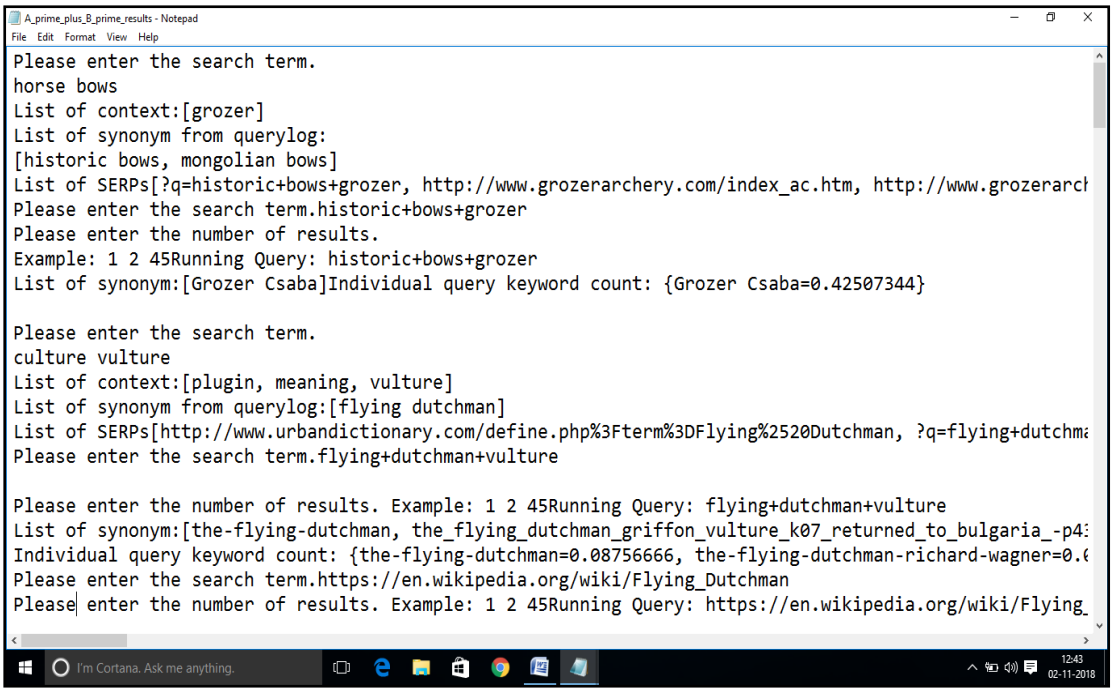


Fig. B.11 More Results of the proposed work to find entity synonyms

Table 4.2 of Chapter IV shows the comparison between the conventional and the proposed approach.

APPENDIX-C

Few snapshots of the data set for concept instance file are shown in Fig. C.1 to C.8. It contains 2.7 million concepts which are collected from 1.68 billion web pages. It contains concepts of worldly facts that human being has formed in their mind. The main reason for using this knowledge source is due to the fact that 85% of the searches contain concepts and/or instances that exist in PROBASE.

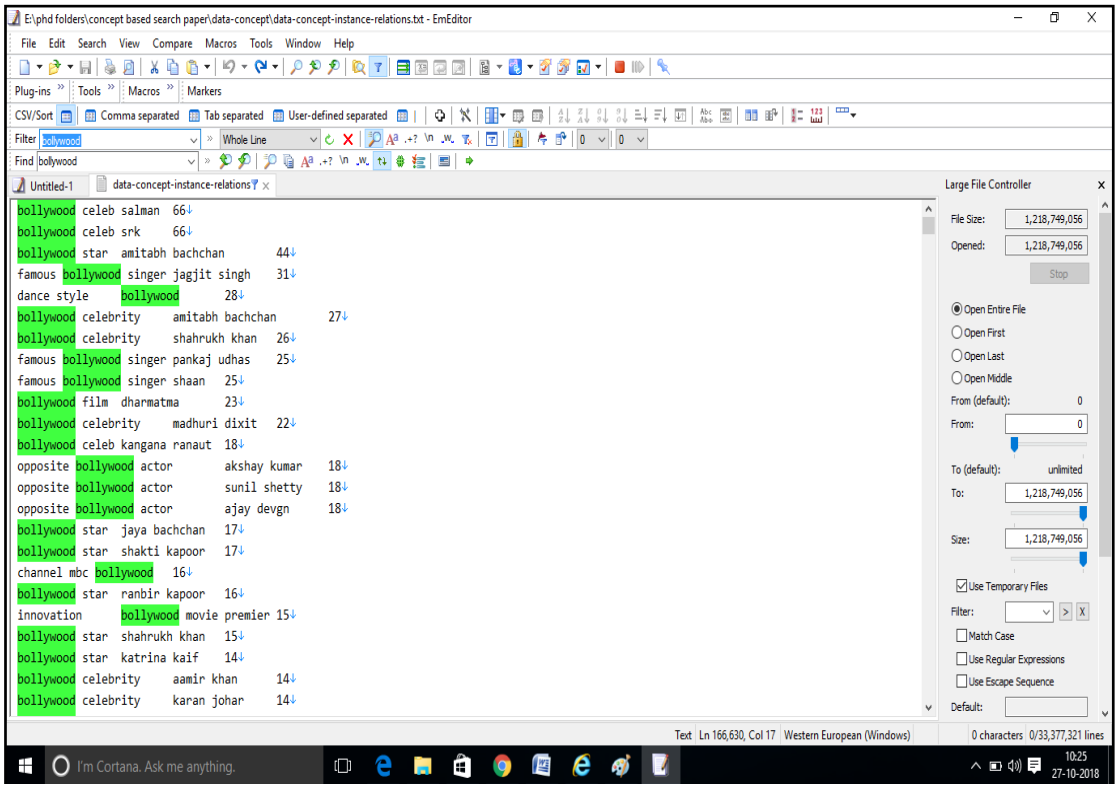


Fig. C.1 Snapshot when the keyword *bollywood* is filtered in knowledge base

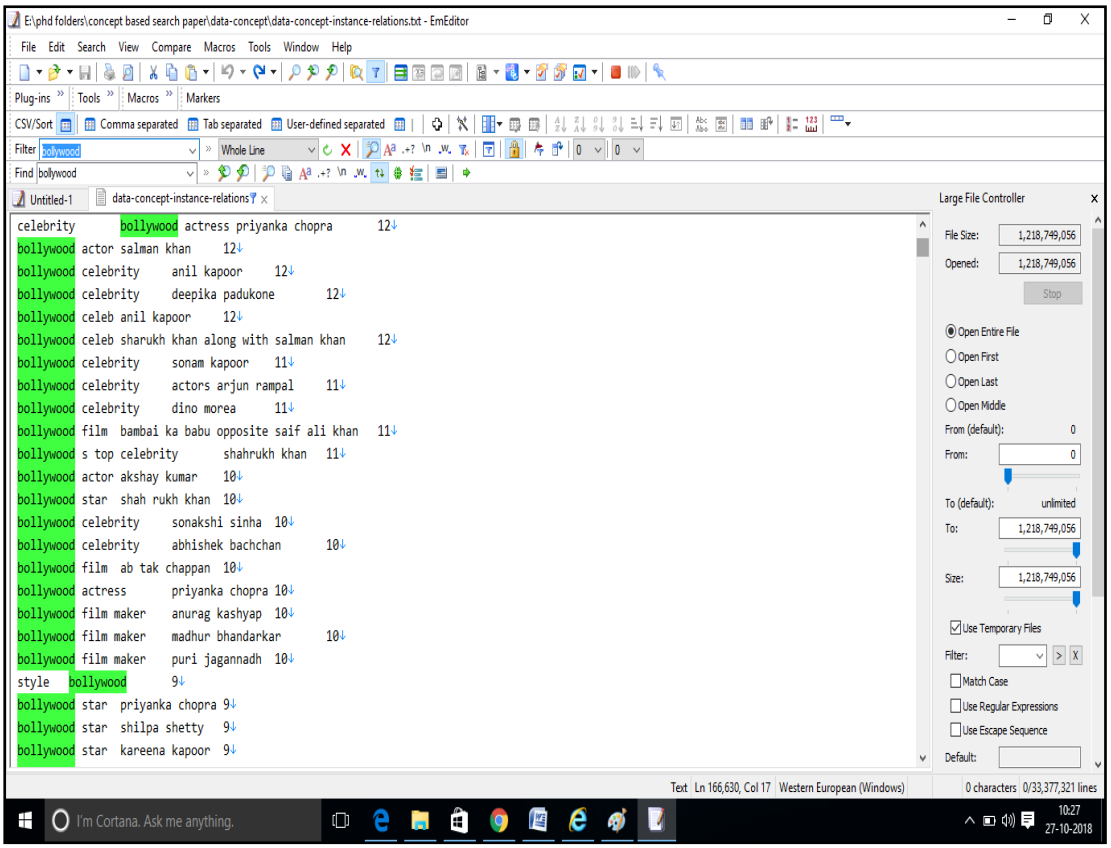


Fig. C.2 Snapshot when the keyword *bollywood* is filtered in knowledge base

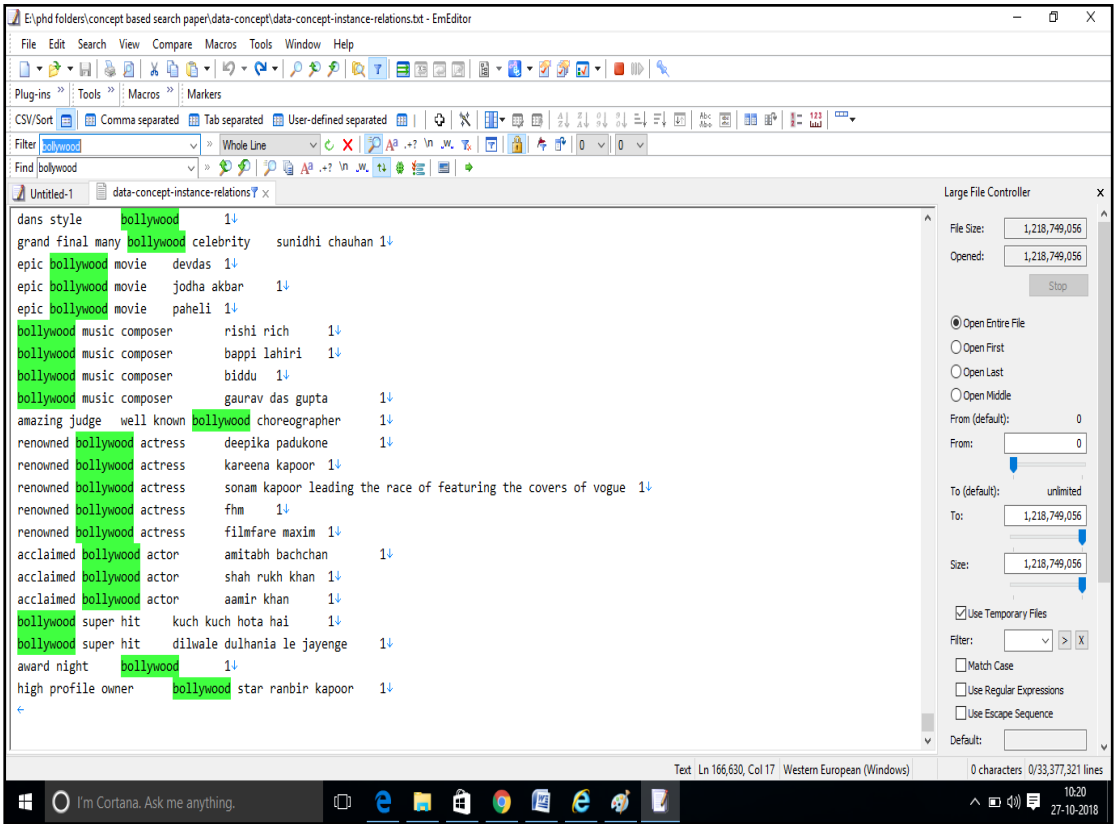


Fig. C.3 Snapshot when the keyword *bollywood* is filtered in knowledge base

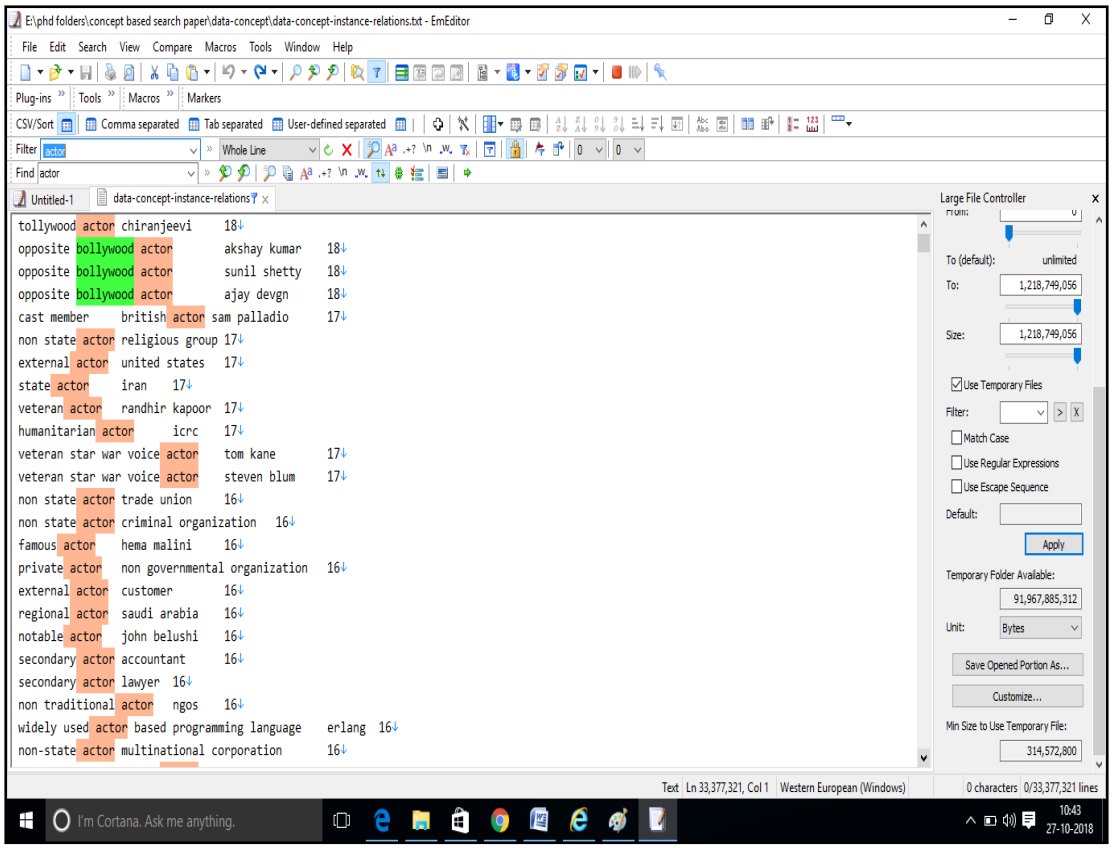


Fig. C.4 Snapshot when the keyword *actor* is filtered in knowledge base

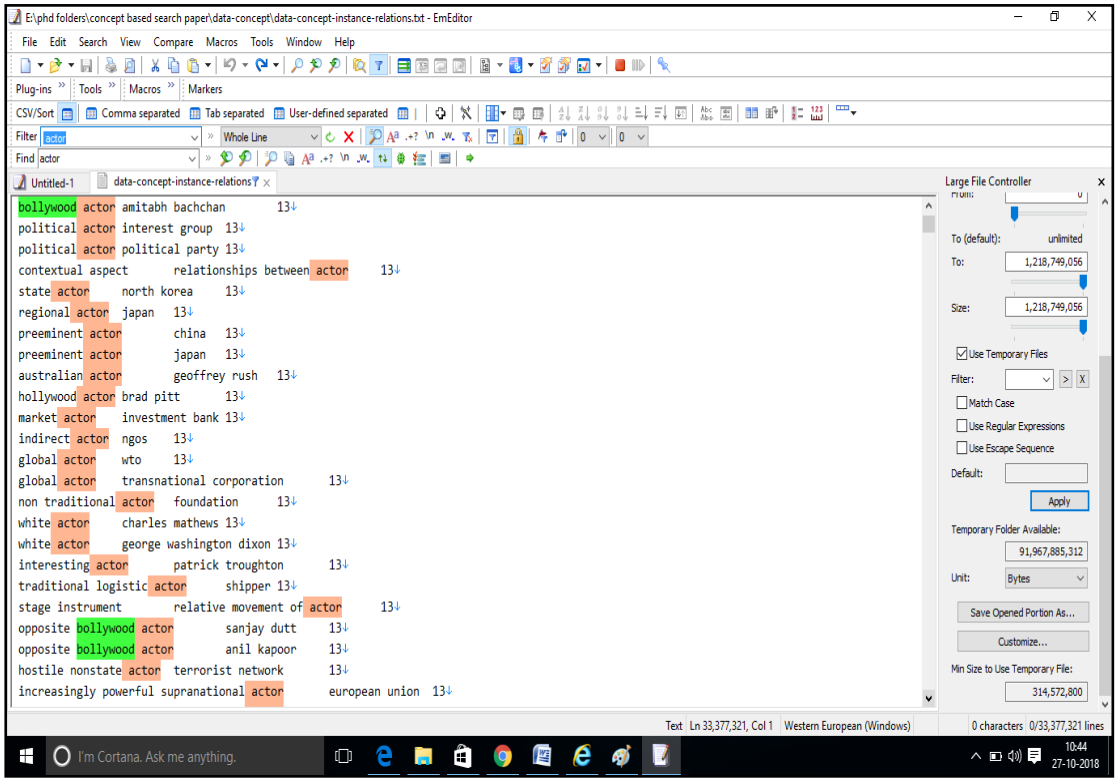


Fig. C.5 Snapshot when the keyword *actor* is filtered in knowledge base

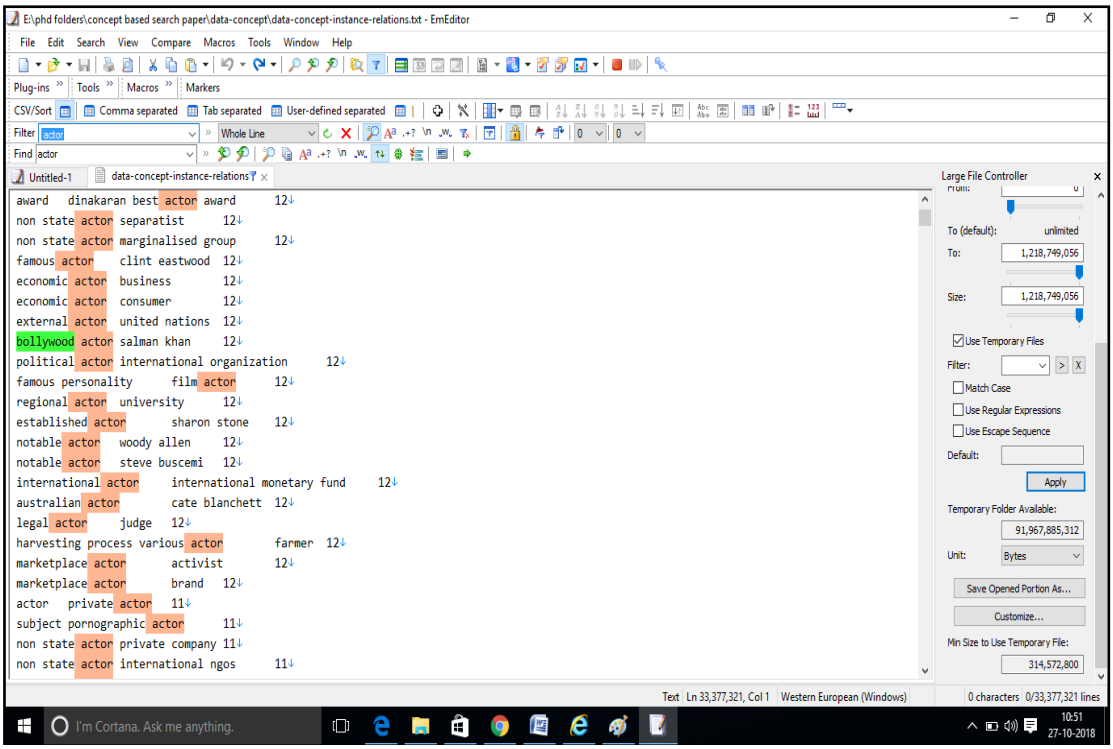


Fig. C.6 Snapshot when the keyword *actor* is filtered in knowledge base

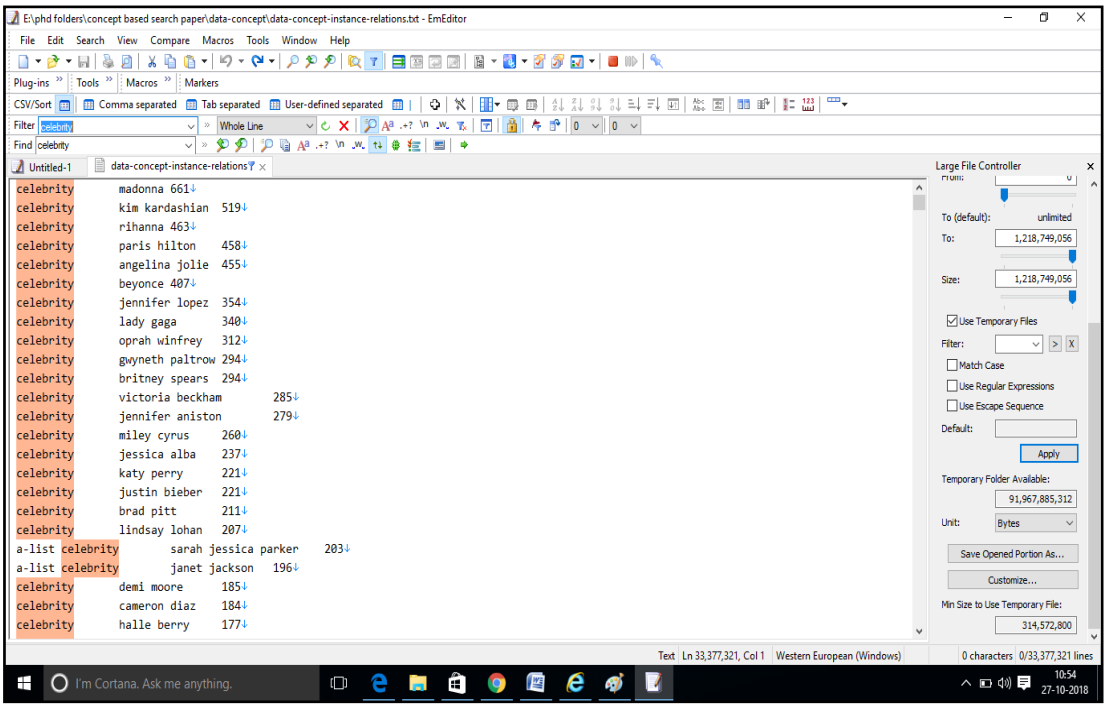


Fig. C.7 Snapshot when the keyword *celebrity* is filtered in knowledge base

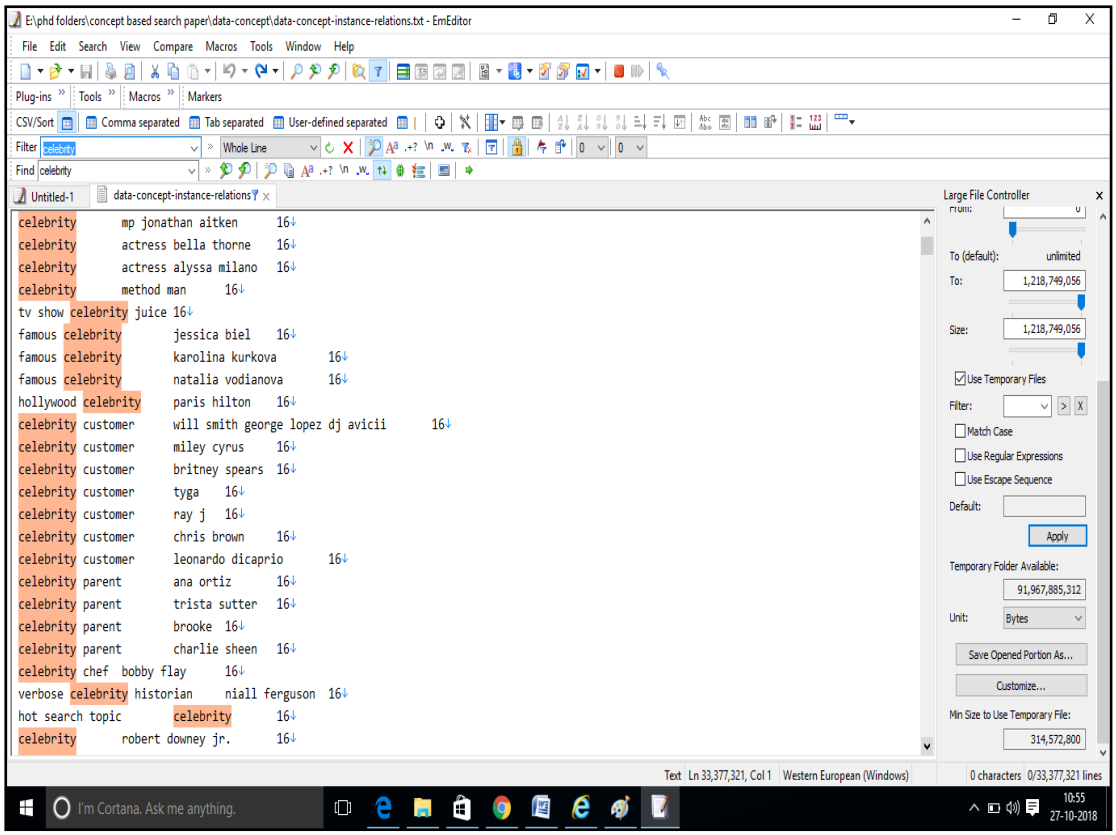


Fig. C.8 Snapshot when the keyword *celebrity* is filtered in knowledge base

The working of the proposed algorithm for the concept actor is shown in Fig. C.1.

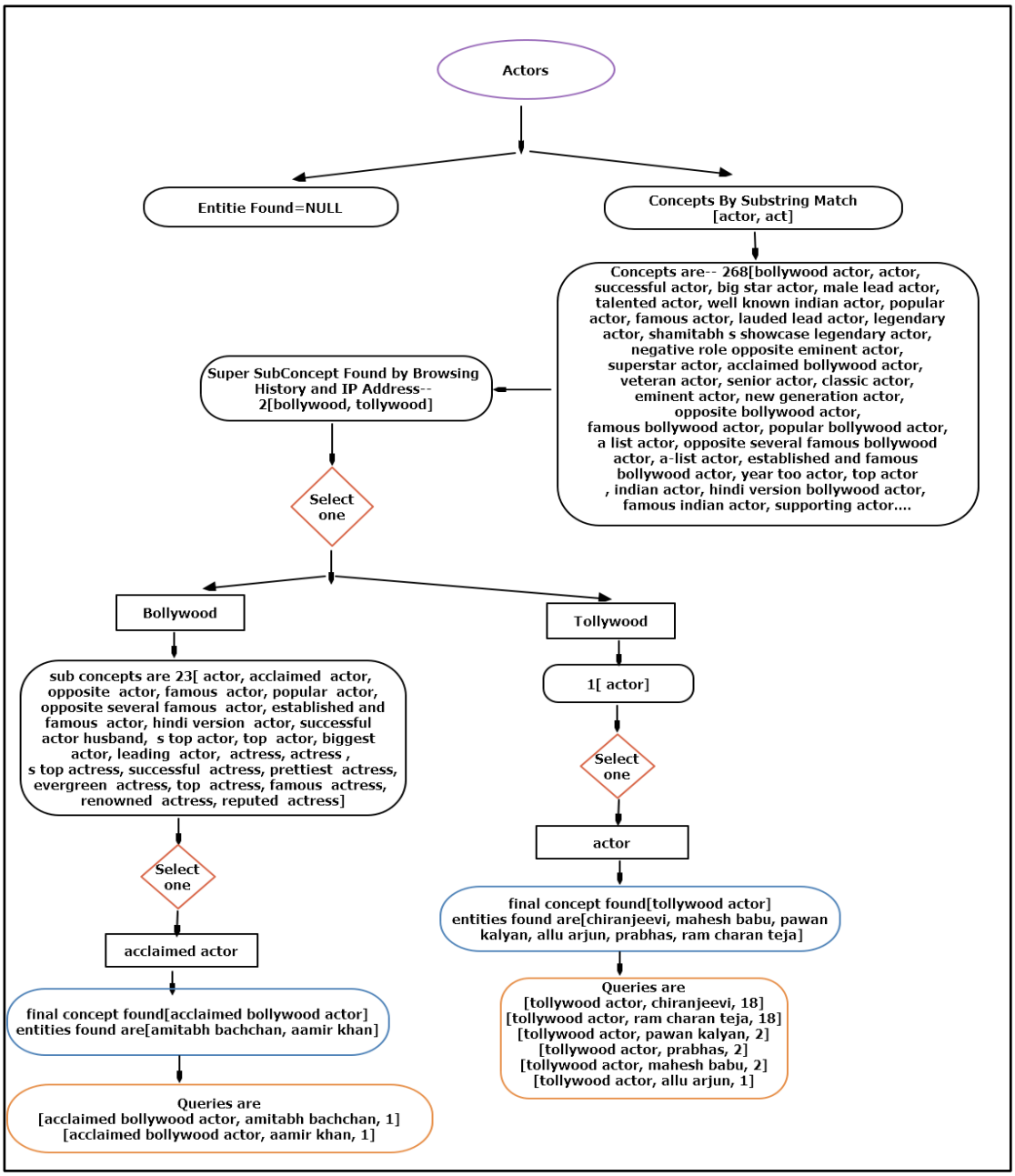
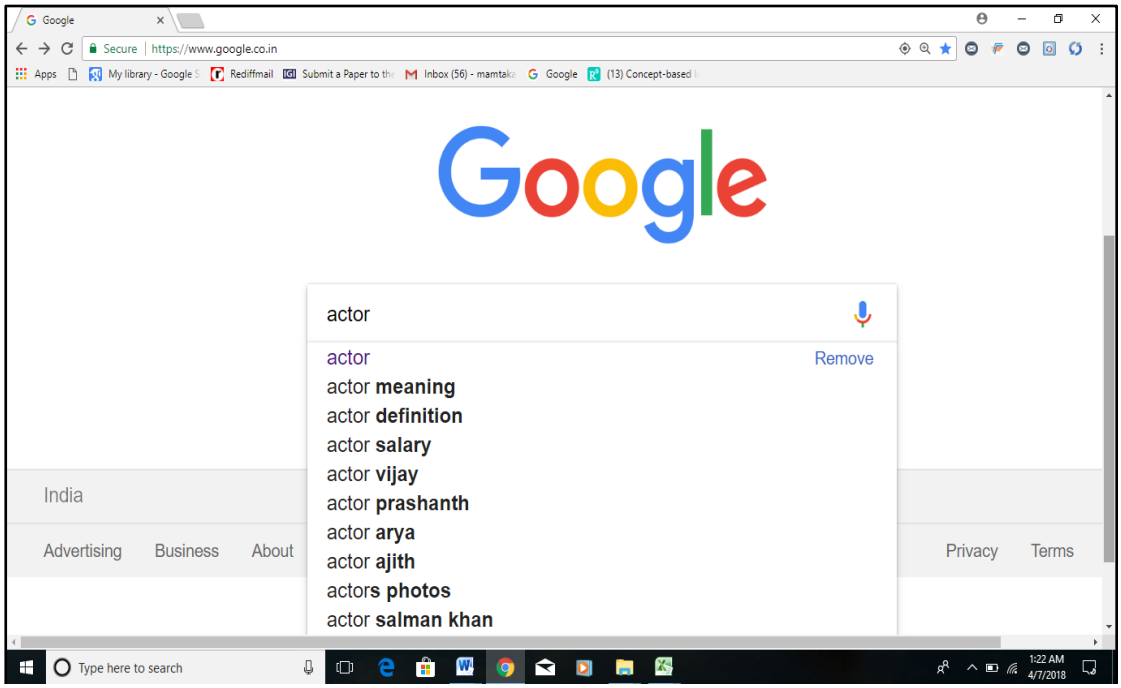


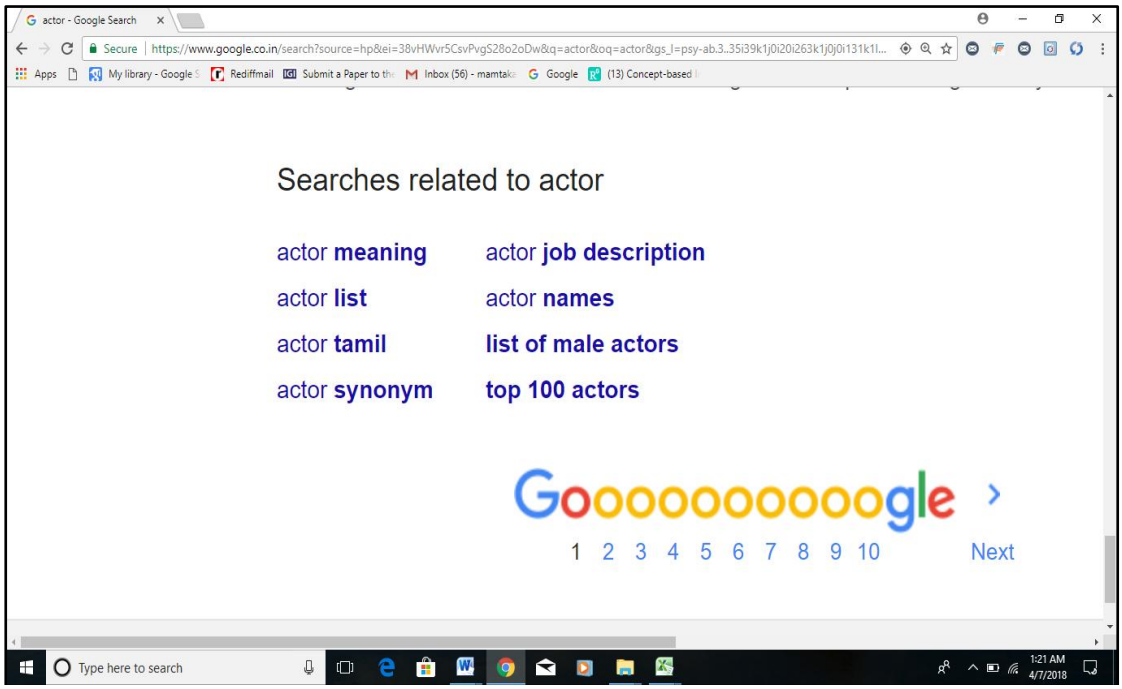
Fig. C.9 Working of the algorithm for the query *Actors*

Query suggestion technique discussed in Section 5.5 of Chapter V is implemented using JAVA eclipse neon. At the front end, a web based project is developed which can be used by anyuser, anywhere to run the implementation of the proposed system. we use PROBASE as a back end tool to extract instances corresponding to the concept. Note that due to privacy concerns we cannot share all the details of the coding. The results of the proposed system has been compared with the topmost

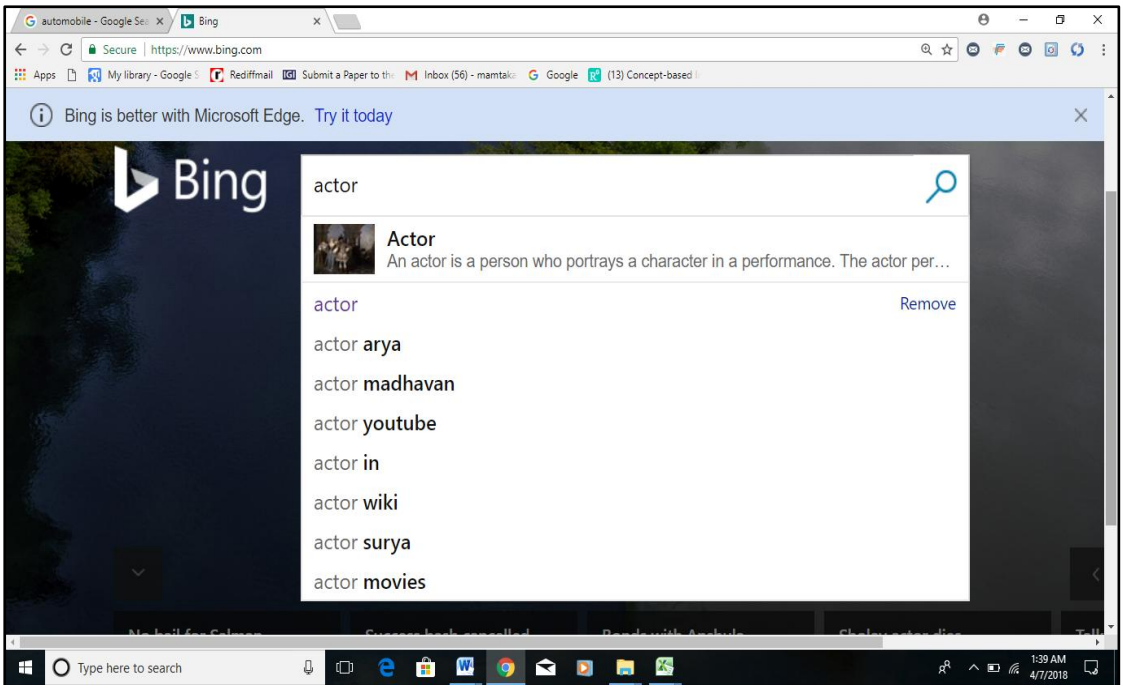
search engines such as Google, Bing and Yahoo.The snapshots of the proposed system for the query *actor* and *bollywood actor* are shown in Fig. C.10- C.23.



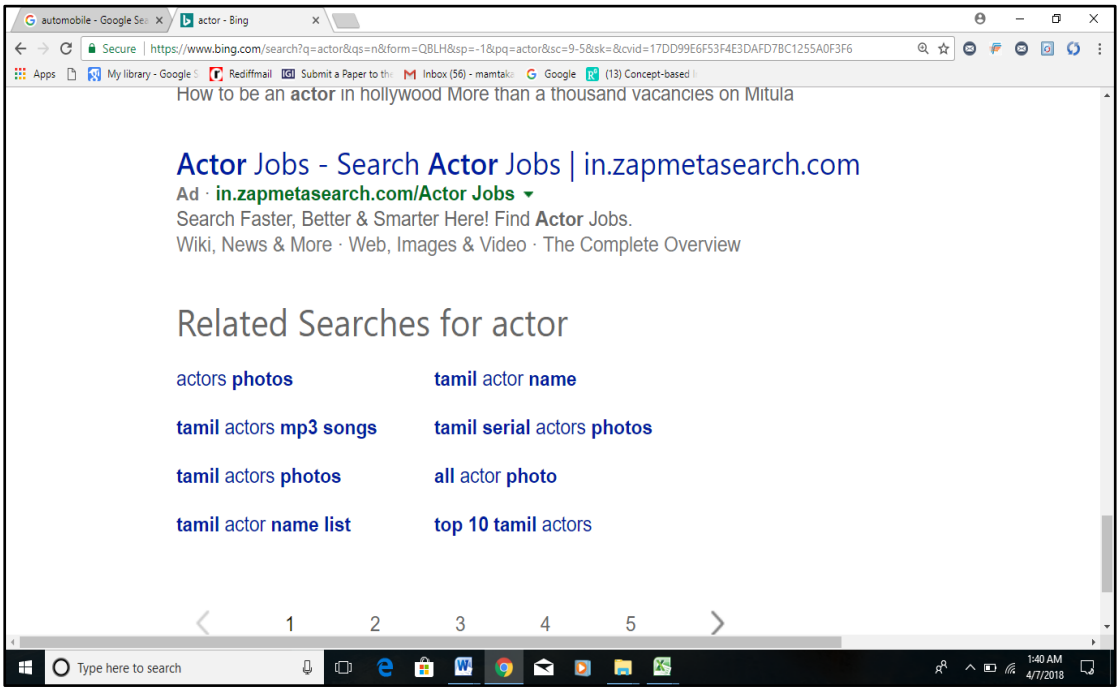
C.10: Query expansion result for the query *actor* by Google search engine



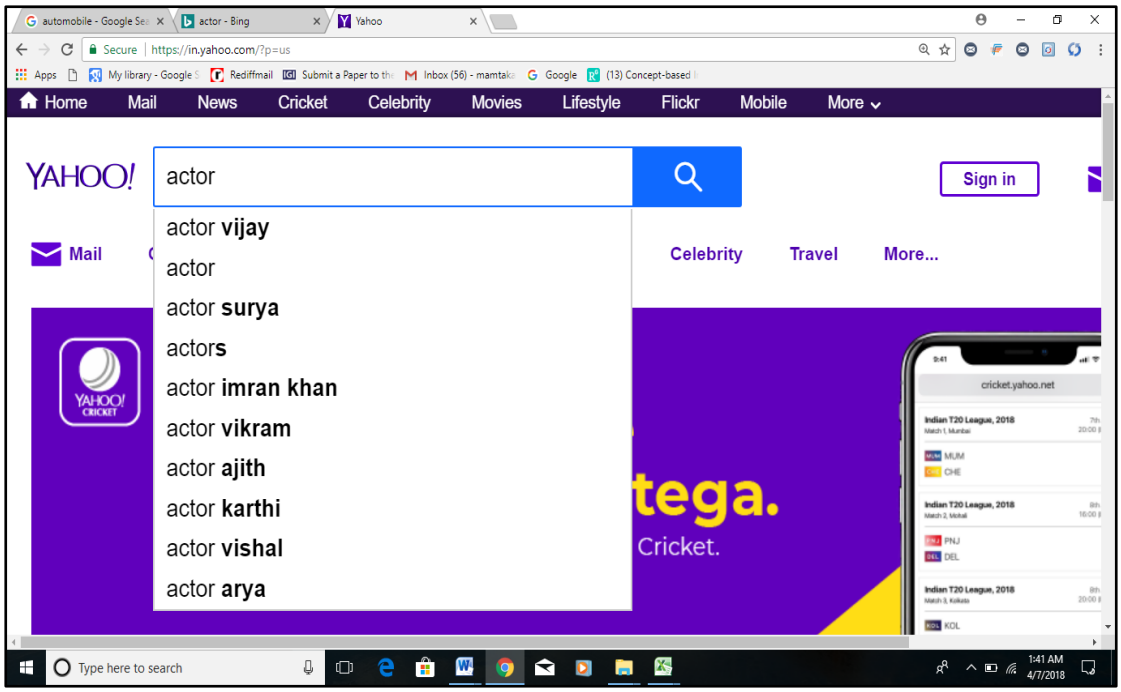
C.11: Query Suggestion result for the query actor by Google search engine



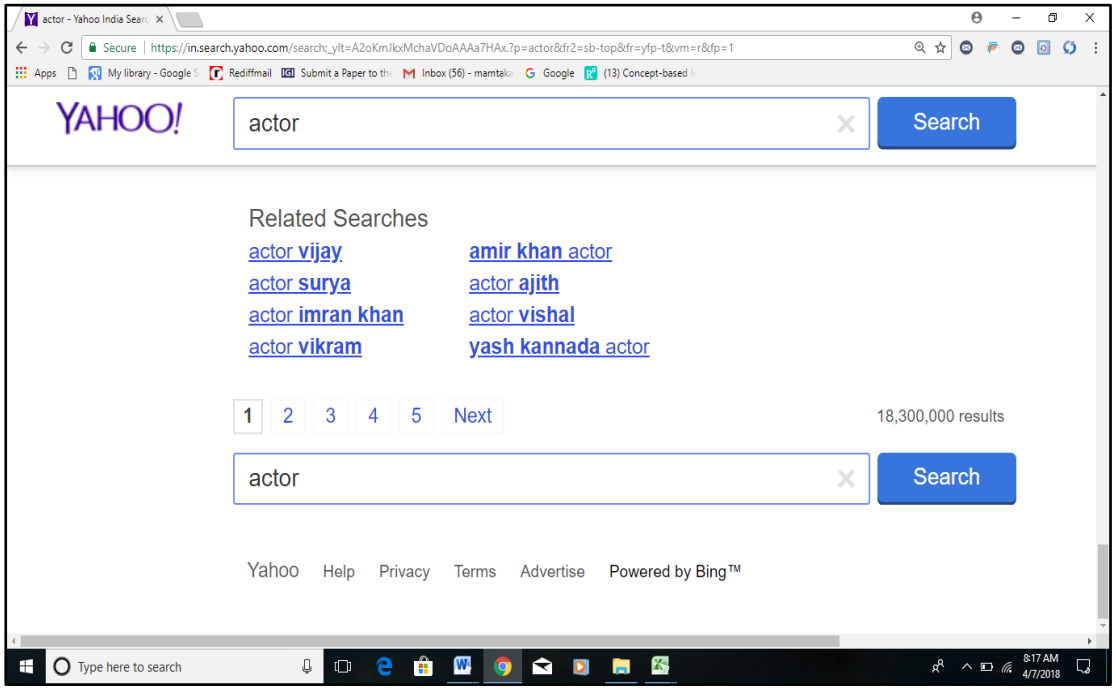
C.12: Query expansion result for the query *actor* by Bing search engine



C.13: Query Suggestion result for the query *actor* by Bing search engine



C.14: Query expansion result for the query *actor* by Yahoo search engine



C.15: Query Suggestion result for the query *actor* by Yahoo search engine

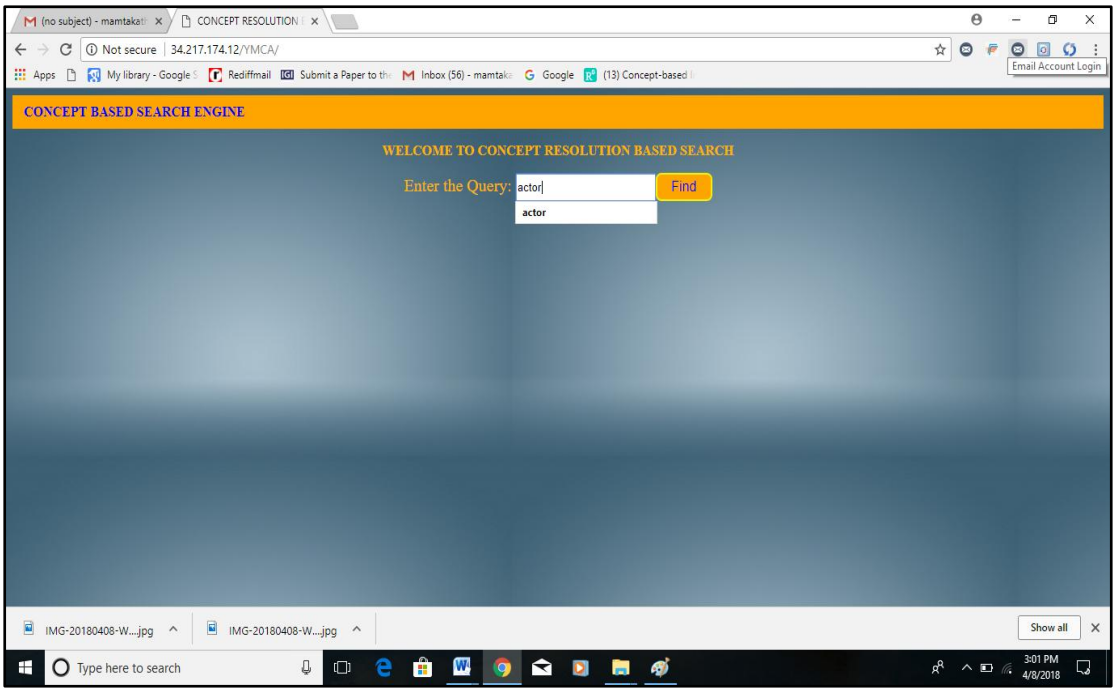
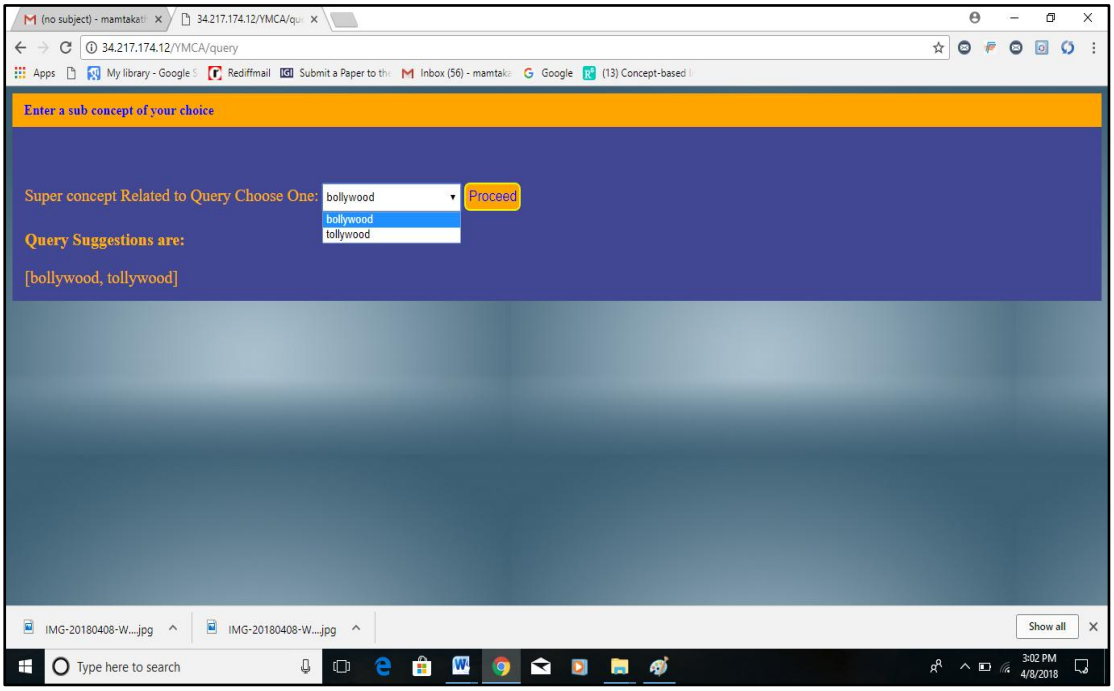
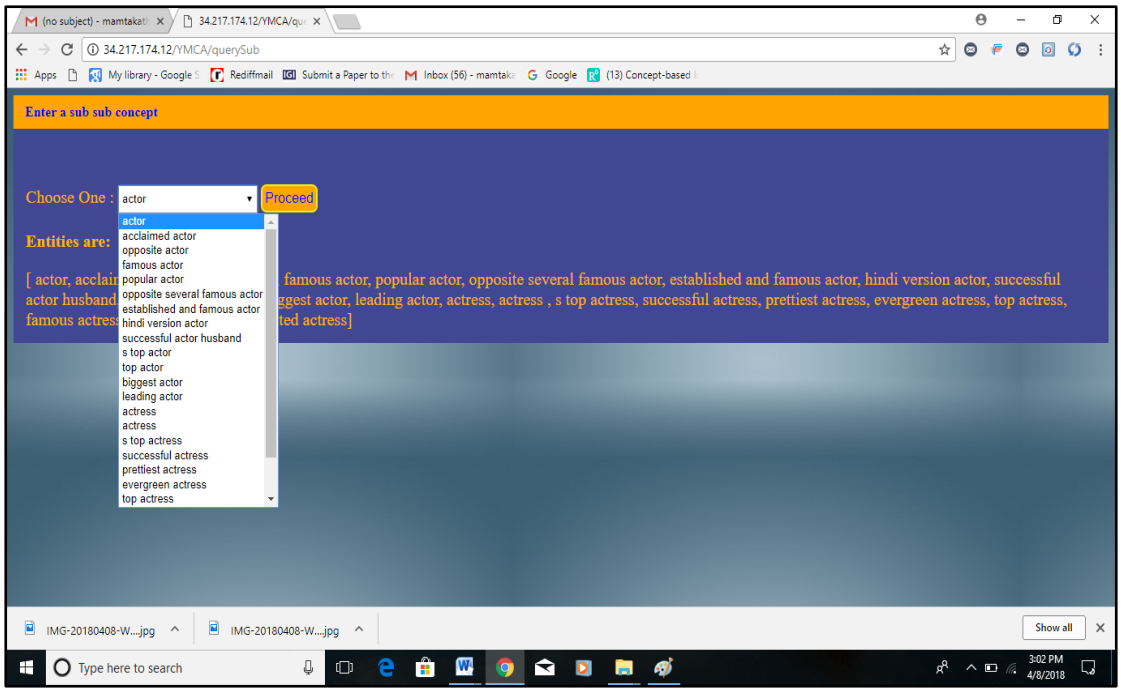


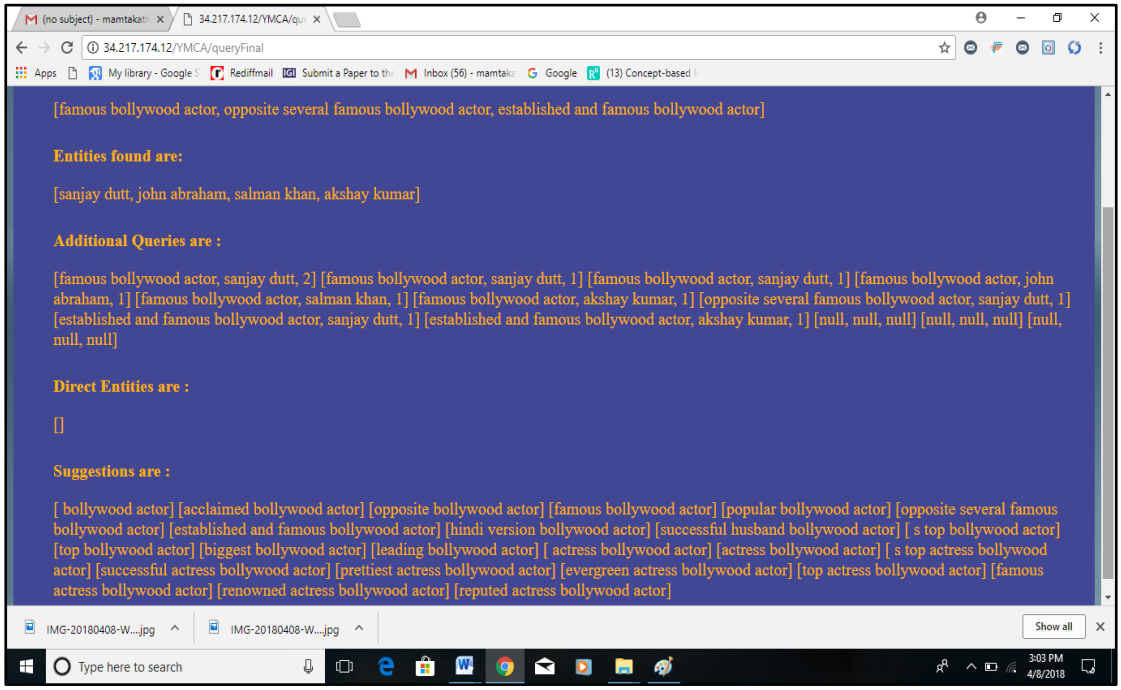
Fig C.16: User interface of Concept Resolution based Search for query *actor*



C.17: First level Sub-concepts related to the query *Actor*



C.18: Second level Sub-concepts related to the input query Actor



C.19: Query suggestion by the proposed system and entities corresponding to the concept *Actor* along with the association

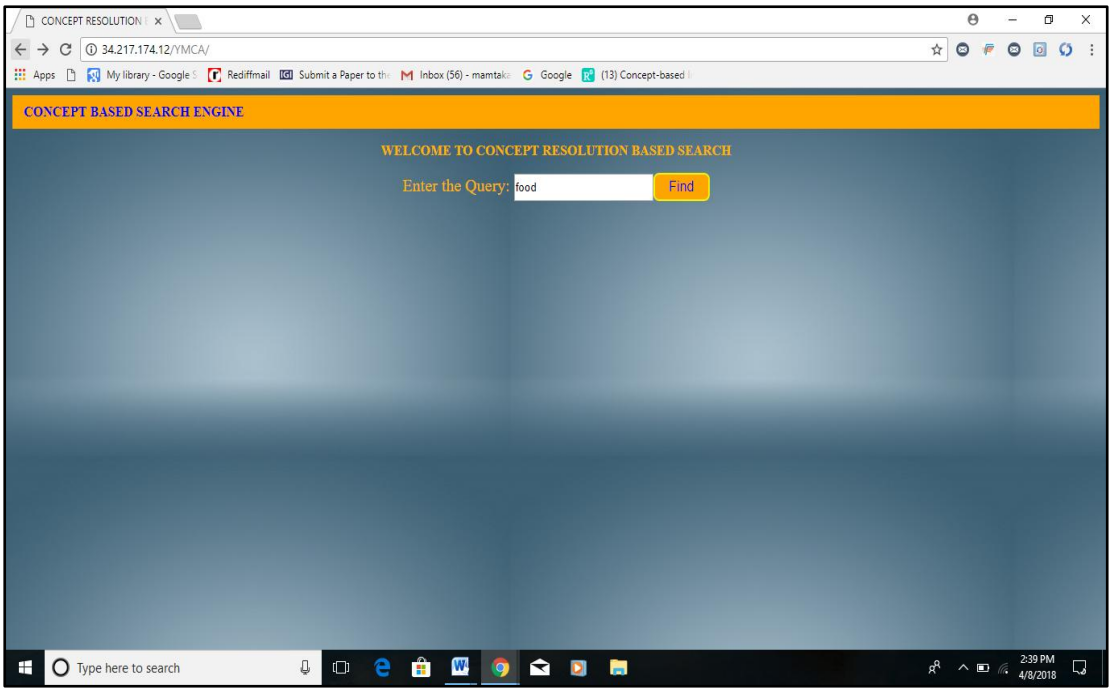
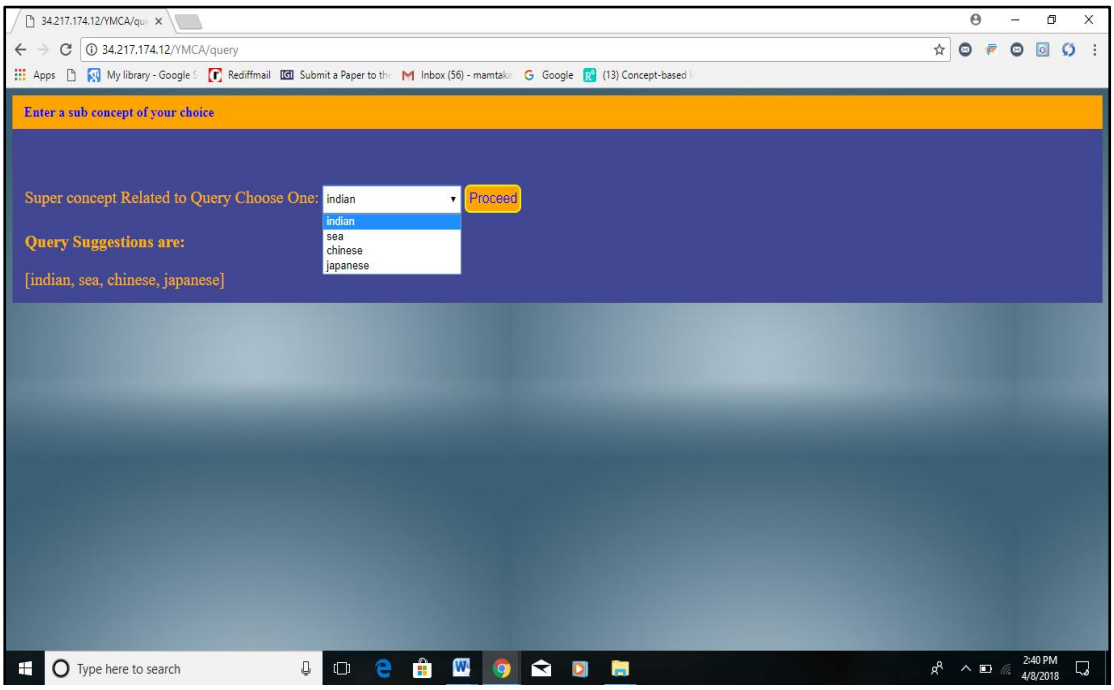
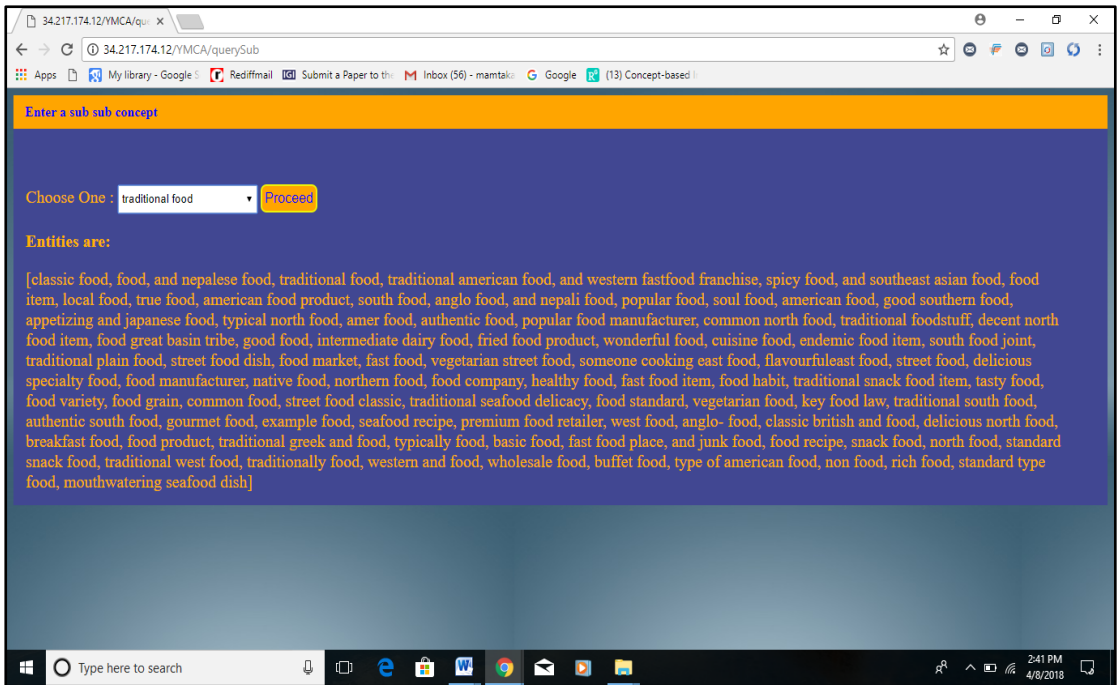


Fig C.20: User interface of Concept Resolution based Search for query *Food*



C.21: First level Sub-concepts related to the query *Food*



C.22: Second level Sub-concepts related to the query *Food*

C.23: Query suggestion by the proposed system and entities corresponding to the concept *Food* along with the association

APPENDIX-D

A fragment of Query Log corresponding to different queries issued by various users in AOL query log are given in Table D.1

Table D.1 Sample of Query Log

Anon ID	Query	QueryTime	Item Rank	ClickURL
144	www.bostonredsox	28-03-2006 18:12	1	http://boston.redsox.mlb.com
144	www. findmassmoney.com	15-04-2006 07:34	1	http://www.findmassmoney.com
144	www.herbchambers.com	23-04-2006 09:23	3	http://www.carsearch.com
144	www.herbchambers.com	23-04-2006 09:23	1	http://www.herbchambers.com
144	www.eastern bank.com	10-05-2006 10:55	1	http://www.easternbank.com
227	psychiatric disorders	02-03-2006 17:30	1	http://www.merck.com
227	psychiatric disorders	02-03-2006 17:30	3	http://allpsych.com
227	cyclothymia	02-03-2006 17:34	1	http://www.psycom.net
227	cyclothymia	02-03-2006 17:34	5	http://www.mental-health-matters.com
227	midwestcenter	02-03-2006 17:35	1	http://www.midwestcenter.com
227	us magazine	05-03-2006 22:00	1	http://www.usmagazine.com
227	areslite	07-03-2006 15:19	2	http://aresgalaxy.sourceforge.net
227	areslite	07-03-2006 22:03	8	http://www.download-free.programas-gratis.net
227	areslite	07-03-2006 22:03	10	http://www.areslite.officialares.com
227	harrisburg pa hotels	08-03-2006 23:50	2	http://www.harrisburgpahotels.worldweb.com
227	section 8 housing pa	09-03-2006 22:54	1	http://www.affordablehousingonline.com
227	acute hepatitus	10-03-2006 14:00	1	http://www.medhelp.org
227	acute hepatitus	10-03-2006 14:00	7	http://www.hc-sc.gc.ca
227	acute hepatitis	10-03-2006 14:03	1	http://www.netdoctor.co.uk
227	pugs	11-03-2006 01:30	1	http://www.pugs.com
227	pugs	11-03-2006 01:30	6	http://www.puppydogweb.com
227	penn national building in harrisburg pa	07-05-2006 00:32	1	http://www.visithhc.com
227	penn national building in harrisburg pa	07-05-2006 00:32	8	http://www.helloharrisburg.com
227	parking garage in harrisburg pa	07-05-2006 00:35	9	http://harrisburg.citysearch.com
227	parking authority in harrisburg pa	07-05-2006 00:37	1	http://hpaparking.homestead.com
227	welfare fraud in harrisburg pa	09-05-2006 22:13	1	http://www.oig.state.pa.us
227	mental disorders	09-05-2006 23:31	3	http://www.mentalhealth.com
227	cyclothymic disorder	09-05-2006 23:32	1	http://www.mentalhealth.com
227	cyclothymic disorder	09-05-2006 23:32	4	http://www.mental-health-matters.com

227	volunteers of america	14-05-2006 22:32	1	http://www.voa.org
227	american spirit tobacco	15-05-2006 22:39	8	http://www.americanspirit-europe.com
227	american spirit tobacco	15-05-2006 22:39	1	http://www.nascigs.com
227	interstate 81	16-05-2006 23:13	1	http://www.interstate-guide.com
227	lebanon pa	16-05-2006 23:19	2	http://www.city-data.com
227	the brook at colonial park	18-05-2006 22:59	2	http://www.peoplewithpets.com
227	the brook at colonial park	18-05-2006 22:59	4	http://www.apartmentguide.com
227	subsidized housing in harrisburg pa	18-05-2006 23:04	3	http://www.tenant.net
227	subsidized housing in harrisburg pa	18-05-2006 23:07	44	http://www.northpennlegal.org
227	subsidized housing in harrisburg pa	18-05-2006 23:07	45	http://www.papartnerships.org
227	subsidized housing in harrisburg pa	18-05-2006 23:08	56	http://www.phil.frb.org
227	www.nik	19-05-2006 23:15		
227	birth control methods	21-05-2006 00:53	2	http://www.plannedparenthood.org
227	lime wire	21-05-2006 01:11	1	http://www.limewire.com
227	hershey medical center	23-05-2006 23:22	1	http://www.hmc.psu.edu
227	williams grove amusement park	23-05-2006 23:55	1	http://www.williamsgrovepark.com
227	mattress warehouse	26-05-2006 00:23	1	http://www.sleephappens.com
227	surplus furniture warehouse	26-05-2006 00:40		
227	furniture factory outlet	26-05-2006 00:42	1	http://www.furniturefactoryoutlet.com
227	furniture factory outlet	26-05-2006 00:42	8	http://www.mariettaoutlet.com
227	eddie's furniture gallery	26-05-2006 00:48	1	http://yellowpages.superpages.com
227	furniture stores in harrisburg pa	26-05-2006 00:49		
227	furniture stores in harrisburg pa	26-05-2006 00:50		
227	mattress stores in harrisburg pa	26-05-2006 00:51	7	http://www.colonialfurniture.com
227	mattress stores in harrisburg pa	26-05-2006 00:51	10	http://harrisburg-pa.stores.premierguide.com
227	xanax	26-05-2006 01:21		
227	car dealers for bad credit	28-05-2006 23:05		
309	www.gamewinners.com	01-03-2006 10:02	1	http://www.gamewinners.com
309	www.gamewinners.com	01-03-2006 11:09	1	http://www.gamewinners.com
309	www.goggle.com	01-03-2006 12:29	3	http://www.goggle.net
309	www.google.com	01-03-2006 12:31	1	http://www.google.com
309	www.pokemon.com	01-03-2006 14:11	2	http://www.pokemon.com
309	www.google.com	01-03-2006 14:23	1	http://www.google.com
309	www.gamewinners.com	01-03-2006 15:20	1	http://www.gamewinners.com
309	www.warofthemonsters.com	01-03-2006 15:50	3	http://www.us.playstation.com
309	map of iraq.com.	02-03-2006 11:40	3	http://www.comcast.net

309	www.gamewinners.com	07-03-2006 15:34	1	http://www.gamewinners.com
309	www.gamewinners	07-03-2006 16:33	1	http://www.gamewinners.com
309	www.gamewinners.com	08-03-2006 15:33	1	http://www.gamewinners.com
309	www.gamewinners.com	08-03-2006 15:33	1	http://www.gamewinners.com
309	www.pokemon.com	10-03-2006 12:35	1	http://www.pokemon.com
309	www.gamewinners.com	10-03-2006 12:47	1	http://www.gamewinners.com
309	www.gamewinners.com	10-03-2006 14:49	1	http://www.gamewinners.com
309	www.amishdonkey.com	10-03-2006 15:09	2	http://www.amishdonkey.com
309	google	18-03-2006 16:59	1	http://www.google.com
309	nobel phone cards.com.	21-03-2006 13:43	7	http://www.linkreferral.com
309	www.my billing.com.	29-03-2006 09:26	3	http://www.socalgas.com
309	www.wunderground.com global dr.html	20-04-2006 07:51	1	http://www.google.com
309	www.pokemon.com	28-04-2006 15:36	1	http://www.pokemon.com
309	www.pokemon.com	28-04-2006 15:44	1	http://www.pokemon.com
309	www.pokemon.com	28-04-2006 15:44	2	http://www.pokemon.com
309	www.pokemon.com	28-04-2006 15:44	1	http://www.pokemon.com
309	www.pokemon.com	05-05-2006 13:22	1	http://www.pokemon.com
309	www.pokemon.com	05-05-2006 13:22	1	http://www.pokemon.com
309	unwritten	05-05-2006 13:39	2	http://www.amazon.co.uk
309	whec tv in rochester ny	11-05-2006 14:54	1	http://www.10nbc.com
309	girls ask	23-05-2006 15:51	1	http://www.gogirlsonly.org
309	www.ebay.com	26-05-2006 19:16	1	http://www.ebay.com
309	www.gamewinners.com	30-05-2006 15:18	1	http://www.gamewinners.com
309	-	31-05-2006 15:44	2	http://www.learningfamily.net
309	pen pals for kids'	31-05-2006 15:50	1	http://www.zen.org
366	intravenous	01-03-2006 17:16	3	http://en.wikipedia.org
647	rabbit hole	01-03-2006 22:11	1	http://www.rabbithole.org
647	rabbit hole the broadway play	01-03-2006 22:15	2	http://www.entertainment-link.com
647	la maganette 825 3rd avenue	01-03-2006 22:29	3	http://www.justsalsa.com
647	betsy johnson	12-03-2006 13:17	1	http://www.betseyjohnson.com
647	lord and taylor	12-03-2006 13:30	6	http://www.victoriana.com
647	maps	15-03-2006 20:52	1	http://www.mapquest.com
647	maps	15-03-2006 20:52	1	http://www.mapquest.com
647	new york times	19-03-2006 20:09	1	http://www.nytimes.com
647	u.s. map	24-03-2006 16:23	1	http://www.mapsonus.com
647	u.s. map	24-03-2006 16:23	3	http://www.infoplease.com

647	university of delaware	24-03-2006 16:35	1	http://www.udel.edu
647	daisy haze	24-03-2006 17:23	1	http://www.grunnenrocks.nl
647	daisy haze	24-03-2006 17:23	6	http://www.fiql.com
647	daisy haze	24-03-2006 17:23	7	http://www.daisyhaze.tk
647	pennstation	25-03-2006 17:58	1	http://www.penn-station.com
647	port authority	25-03-2006 18:05	1	http://www.panynj.gov
647	penn station trains	25-03-2006 18:09	1	http://www.njtransit.com
647	coldstone	26-03-2006 21:29	1	http://www.coldstonecreamery.com
647	balloon bouquets	26-03-2006 21:29	1	http://www.balloonbouquets.com
647	balloon bouquets new york	26-03-2006 21:33	1	http://www.balloonbouquetsnyc.com
647	four faced pub	01-04-2006 11:46	1	http://www.thefour-facedliar.com
647	cliff notes	02-04-2006 18:32	1	http://www.cliffsnotes.com
647	cliff notes	02-04-2006 18:32	3	http://www.antistudy.com
647	cliff notes	03-04-2006 22:23	1	http://www.cliffsnotes.com
647	lirr	12-05-2006 18:51	3	http://www.lirr.org
706	scarlet fever	01-03-2006 14:48	1	http://kidshealth.org
706	online casino	03-03-2006 18:42	1	http://www.goldenpalace.com
706	lotto	11-03-2006 06:58	2	http://www.texaslotto.com
706	lotto	11-03-2006 06:58	1	http://www.txlottery.org
706	lotto	16-03-2006 09:30	1	http://www.txlottery.org
706	lotto	18-03-2006 16:39	2	http://texaslotto.com
706	lotto	18-03-2006 16:39	1	http://www.txlottery.org
706	lotto	19-03-2006 08:12	1	http://www.txlottery.org
706	casino	19-03-2006 13:06	1	http://www.casino.com
706	on line casino	19-03-2006 14:29	1	http://www.goldenpalace.com
706	dvdmovies	19-03-2006 15:25	1	http://www.dvdmovies.com
706	mickey dolenz	25-03-2006 18:24	1	http://www.mickydolenz.com
706	mickey dolenz	25-03-2006 18:24	2	http://www.mickydolenz.com
706	mickey dolenz	25-03-2006 18:24	10	http://www.familyfirst.com
706	peter tork	25-03-2006 20:20	1	http://www.petertork.com
706	peter tork	25-03-2006 20:20	3	http://www.monkees.net
706	peter tork	25-03-2006 20:20	2	http://www.imdb.com
706	peter tork	25-03-2006 20:20	4	http://www.monkees.net
706	peter tork	25-03-2006 20:20	4	http://www.monkees.net
706	mike nesmith	25-03-2006 22:08	4	http://www.nezfriends.com
706	davy jones	25-03-2006 22:14	3	http://www.imdb.com

706	davy jones	25-03-2006 22:14	1	http://www.davyjones.net
706	randy scouse git	27-03-2006 11:23	1	http://colli.tripod.com
706	randy scouse git	27-03-2006 11:23	2	http://www.monkees.net
706	randy scouse git	27-03-2006 11:23	4	http://www.amiright.com
706	mickey dolenz	27-03-2006 11:29	4	http://www.amdest.com
706	this just doesnt seem to be my day	28-03-2006 16:17	2	http://www.lyricsdownload.com
706	sweet young thing	28-03-2006 16:23	1	http://www.lyricsdepot.com
706	sweet young thing	28-03-2006 16:23	2	http://www.monkees.net
706	lotto	14-04-2006 09:05	4	http://www.txlottery.org
706	lotto	23-04-2006 07:23	4	http://www.txlottery.org
706	lotto	23-04-2006 07:23	4	http://www.txlottery.org
706	laberge casino	23-04-2006 15:30	1	http://casino777.in.ua
706	laberge casino	23-04-2006 15:30	2	http://casino777.in.ua
706	laberge casino	23-04-2006 15:30	3	http://www.ramada.com
706	marathon key	24-04-2006 08:55	1	http://www.fla-keys.com
706	asteroid	27-04-2006 21:04	1	http://impact.arc.nasa.gov
706	lotto	29-04-2006 09:44	4	http://www.txlottery.org
706	bill hickock	29-04-2006 12:30	1	http://www.abacom.com
706	lotto	30-04-2006 12:07	4	http://www.txlottery.org
706	erich von manstein	30-04-2006 14:05	1	http://www.spartacus.schoolnet.co.uk
706	lotto	02-05-2006 08:56	4	http://www.txlottery.org
706	shrubbury	10-05-2006 15:00	2	http://www.allthetests.com
706	tattoos	12-05-2006 13:38		
706	buspirone	18-05-2006 15:45	1	http://www.mentalhealth.com
706	buspirone	18-05-2006 15:45	2	http://www.rxlist.com
706	northwestern university	18-05-2006 16:11	1	http://www.northwestern.edu
808	yahoo	01-03-2006 02:14	1	http://www.yahoo.com
808	mapquest com	04-05-2006 21:54	1	http://www.mapquest.com
808	mapquest com	04-05-2006 21:54	1	http://www.mapquest.com
808	mapquest com	06-05-2006 15:22	1	http://www.mapquest.com
808	-	19-05-2006 20:14	1	http://www.google.com
808	cartoon network	29-05-2006 22:31	2	http://www.cartoonnetwork.co.uk
808	arcadepod	29-05-2006 22:32	1	http://www.arcadepod.com
1038	sean fraser	02-03-2006 11:30	2	http://gotigersgo.collegesports.com
1038	wait till u see my	02-03-2006 16:40	4	http://profile.myspace.com
1038	dvd 9	03-03-2006 16:26	2	http://www.afterdawn.com

1038	www.ghanaweb.com	03-03-2006 16:33	1	http://www.ghanaweb.com
1038	shane mcfaul	03-03-2006 16:53	3	http://en.wikipedia.org
1038	shane mcfaul	03-03-2006 16:53	4	http://www.answers.com
1038	shane mcfaul	03-03-2006 16:53	8	http://www.1862.net
1038	shane mcfaul	03-03-2006 16:53	10	http://www.bbc.co.uk
1038	shane mcfaul	03-03-2006 17:27	3	http://www.bebo.com
1038	shane mcfaul	03-03-2006 17:27	8	http://www.football.co.uk
1038	shane mcfaul	03-03-2006 17:27	11	http://www.bbc.co.uk
1038	shane mcfaul	03-03-2006 17:30	14	http://www.uefa.com
1038	shane mcfaul	03-03-2006 17:52	2	http://archives.tcm.ie
1038	shane mcfaul	03-03-2006 17:56	4	http://www.nottscountyfc.premiumtv.co.uk
1038	liam george	03-03-2006 17:57	1	http://www.lutonfc.com
1038	liam george	03-03-2006 17:57	2	http://www.lutonfc.com
1038	liam george	03-03-2006 17:57	3	http://www.kickinmagazine.ie
1038	liam george	03-03-2006 18:01	4	http://www.thisisyork.co.uk
1038	maurice hughes	03-03-2006 18:02		
1038	barnes homer mathew	04-03-2006 20:30	1	http://soccernet.espn.go.com
1038	barnes homer mathew	04-03-2006 20:30	2	http://news.bbc.co.uk
1038	barnes homer mathew	04-03-2006 20:30	2	http://news.bbc.co.uk
1038	scroggins leepaul	04-03-2006 20:33	1	http://www.ncaasports.com
1038	barnes homer mathew	04-03-2006 20:37	1	http://soccernet.espn.go.com
1038	health sector jobs	04-03-2006 22:50	7	http://www.marketwatch.com
1038	health sector jobs	04-03-2006 22:50	10	http://www.prospects.ac.uk
1038	herzing college atlanta	04-03-2006 23:04	1	http://www.herzing.edu
1038	study guide for the foreign service written examination	04-03-2006 23:07	1	http://careers.state.gov
1038	joe afful	05-03-2006 02:52	4	http://www.uslfans.com
1038	joe afful	05-03-2006 02:52	6	http://www.northeastconference.org
1038	joe afful	05-03-2006 02:53	7	http://www.ajc.com
1038	joe afful	05-03-2006 02:53	8	http://www.ajc.com
1038	joe afful	05-03-2006 02:55	2	http://seattletimes.nwsource.com
1038	joe afful	05-03-2006 02:55	9	http://aupanthers.collegesports.com
1038	joe afful	05-03-2006 02:56	8	http://www.fiusports.com
1038	joe afful	05-03-2006 02:56	14	http://goprincetontigers.collegesports.com
1038	psv	05-03-2006 12:05	1	http://english.psv.nl
1038	lee nguyen	05-03-2006 12:06	6	http://soccernet.espn.go.com
1038	brochures for business	05-03-2006 12:10	5	http://www.hp.com

1038	brochures for business	05-03-2006 12:10	6	http://www.hansonmarketing.com
1038	brochures for business	05-03-2006 12:10	8	http://www.smallbusinessbrief.com
1038	brochures for business	05-03-2006 12:10	10	http://www.quickbrochures.com
1038	brochures for business	05-03-2006 12:10	9	http://www.smallbusinessbrief.com
1038	brochures for business	05-03-2006 12:10	7	http://www.printingforless.com
1038	brochure	05-03-2006 12:38	2	http://www.brochure-design.com
1038	brochure	05-03-2006 12:38	6	http://www.createchange.org
1038	brochure	05-03-2006 12:38	10	http://www.acsm.net
1038	baba armando	05-03-2006 12:54	5	http://allafrica.com
1038	baba armando	05-03-2006 12:54	4	http://www.ghanaweb.com
1038	baba armando	05-03-2006 12:54	9	http://highwaycentre.com
1038	baba armando	05-03-2006 12:56	1	http://www.goal.com
1038	how to write fractions in microsoft word	05-03-2006 19:31	1	http://www.ele.uri.edu
1038	how to write fractions in microsoft word	05-03-2006 19:31	2	http://www.powertolearn.com
1038	how to write fractions in microsoft word	05-03-2006 19:40	4	http://www.microsoft.com
1038	how to write fractions in microsoft word	05-03-2006 19:40	1	http://www.ele.uri.edu
1038	how to write fractions in microsoft word	05-03-2006 19:40	1	http://www.ele.uri.edu
1038	dobbs ferry	05-03-2006 21:16	1	http://www.dobbsferry.com
1038	richmond afful	06-03-2006 10:48	1	http://student.plattsburgh.edu
1038	richmond afful	06-03-2006 10:48	2	http://student.plattsburgh.edu
1038	richmond afful	06-03-2006 11:00	5	http://digital-diva.smugmug.com
1038	richmond afful	06-03-2006 11:03	10	http://uslpro.uslsoccer.com
1038	richmond afful	06-03-2006 11:06	12	http://www.psal.org
1038	optygen soccer	06-03-2006 22:08	1	http://www.firstendurance.com
1038	optygen soccer	06-03-2006 22:08	4	http://www.athletes.com
1038	optygen soccer	06-03-2006 22:08	5	http://www.athletes.com
1038	optygen	07-03-2006 19:17	1	http://www.prolithic.com
1038	optygen	07-03-2006 19:17	2	http://www.bodybuilding.com
1038	optygen	07-03-2006 19:17	3	http://www.bodybuilding.com
1038	optygen	07-03-2006 19:17	4	http://www.personalbestnutrition.com
1038	optygen	07-03-2006 19:17	8	http://www.worldcycling.com
1038	optygen	07-03-2006 19:26	5	http://www.naturalstandard.com
1038	optygen	07-03-2006 19:27	3	http://www.ffnmag.com
1038	optygen	07-03-2006 19:28	9	http://www.performancebike.com
1038	optygen	07-03-2006 19:28	10	http://wholeathlete.com
1038	optygen	07-03-2006 19:28	12	http://health-beauty.listings.ebay.com

1038	shalrie joseph	07-03-2006 19:51	1	http://www.revolutionsoccer.net
1038	shalrie joseph	07-03-2006 19:51	1	http://www.revolutionsoccer.net
1038	shalrie joseph	07-03-2006 19:51	5	http://www.caribbeannetnews.com
1038	how to take optygen	08-03-2006 09:36	1	http://www.firstendurance.com
1038	how to take optygen	08-03-2006 09:36	2	http://www.firstendurance.com
1038	how to take optygen	08-03-2006 09:36	3	http://www.pureendurance.net
1038	how to take optygen	08-03-2006 09:36	4	http://www.pureendurance.net
1038	free porn movies	08-03-2006 10:40	1	http://www.timsmovies.com
1038	free porn movies	08-03-2006 10:40	5	http://www.smutgremlins.com
1038	low kupono	08-03-2006 23:41	1	http://www.sligorovers.com
1038	low kupono	08-03-2006 23:41	2	http://www.sligorovers.com
1038	low kupono	08-03-2006 23:42	3	http://www.matchnight.com
1038	low kupono	08-03-2006 23:42	12	http://www.uslsoccer.com
1038	low kupono	08-03-2006 23:42	13	http://www.uslsoccer.com
1038	low kupono	08-03-2006 23:47	1	http://home.hamptonroads.com
1038	low kupono	08-03-2006 23:47	1	http://home.hamptonroads.com
1038	low kupono	08-03-2006 23:47	2	http://www.fansonly.com
1038	low kupono	08-03-2006 23:47	3	http://www.fansonly.com
1038	low kupono	08-03-2006 23:47	4	http://www.montrealimpact.com
1038	low kupono	08-03-2006 23:47	8	http://www.dundalkfc.com
1038	low kupono	08-03-2006 23:47	10	http://usfdons.collegesports.com
1038	low kupono	08-03-2006 23:47	12	http://gohuskies.collegesports.com
1038	low kupono	08-03-2006 23:47	14	http://www.irishfootballonline.com
1038	omar jarun	09-03-2006 16:39	7	http://www.usldiscussions.com
1038	omar jarun	09-03-2006 16:39	10	http://www.mysoccer.com
1038	omar jarun	09-03-2006 16:40	1	http://www.flyernews.com
1038	omar jarun	09-03-2006 16:41	3	http://www.fansonly.com
1038	omar jarun	09-03-2006 16:41	5	http://gotigersgo.collegesports.com
1038	omar jarun	09-03-2006 16:41	7	http://www.usldiscussions.com
1038	harchester	09-03-2006 21:46	3	http://www.harchester.tv
1038	harchester	09-03-2006 21:46	1	http://www.harchester.net
1038	harchester	09-03-2006 21:46	6	http://www.fmunderground.net
1038	harchester	09-03-2006 21:46	7	http://www.premiershirts.net
1038	max bretos	09-03-2006 22:37	1	http://www.soccerloop.com
1038	rentals in atlanta	09-03-2006 23:29	1	http://www.homerentalads.com
1038	describe three strengths skills or personality traits that would	10-03-2006 09:28	2	http://www.vuw.ac.nz

	contribute to your success in the field			
1038	describe three strengths skills or personality traits that would contribute to your success in the field	10-03-2006 09:28	1	http://www.vuw.ac.nz
1038	describe three strengths skills or personality traits that would contribute to your success in the field	10-03-2006 09:28	4	http://www.haas.berkeley.edu
1038	describe three strengths skills or personality traits that would contribute to your success in the field	10-03-2006 09:28	10	http://www.colby-sawyer.edu
1038	describe one weakness or area of personal growth that you feel would represent a challenge to you in the field	10-03-2006 09:31	1	http://dothr.ost.dot.gov
1038	describe one weakness or area of personal growth that you feel would represent a challenge to you in the field	10-03-2006 09:31	7	http://icc3.ucdavis.edu
1038	describe what role sport and or play have had in your life.	10-03-2006 09:42	4	http://www.spiderzrule.com
1038	sample project planning	10-03-2006 09:58	1	http://www.dof.ca.gov
1038	sample project planning	10-03-2006 09:58	5	http://www.themarketingprocessco.com
1038	sebastian svard	10-03-2006 16:12	1	http://www.sebastiansvard.info
1038	sebastian svard	10-03-2006 16:12	2	http://www.soccerbase.com
1038	greyhound	10-03-2006 17:32	1	http://www.greyhound.com
1038	greyhound	10-03-2006 17:32	2	http://www.greyhound.com.au
1038	optygen soccer	11-03-2006 00:58	3	http://www.athletes.com
1038	optygen soccer	11-03-2006 00:58	10	http://www.bizrate.com
1038	optygen soccer	11-03-2006 00:59	1	http://www.ms-se.com
1038	optygen soccer	11-03-2006 00:59	7	http://blog.myspace.com
1038	optygen soccer	11-03-2006 00:59	9	http://hammerheadracing.org
1038	optygen soccer	11-03-2006 00:59	13	http://www.the-best-sporting-goods.com
1038	mario rodriguez	11-03-2006 15:13	9	http://en.wikipedia.org
1038	ibrahim attiku	11-03-2006 23:51	3	http://www.ghanaweb.com
1038	scott schweitzer	12-03-2006 00:12	1	http://charlotte49ers.collegesports.com
1038	scott schweitzer	12-03-2006 00:12	2	http://www.geocities.com
1038	scott schweitzer	12-03-2006 00:12	8	http://www.highbeam.com
1038	scott schweitzer	12-03-2006 00:13	14	http://www.uslfans.com
1038	joe afful	12-03-2006 00:15	3	http://digital-diva.smugmug.com
1038	joseph afful	12-03-2006 00:17	12	http://goprincetontigers.collegesports.com
1038	joe afful	12-03-2006 00:18	6	http://seattlepitch.tripod.com
1038	joe afful	12-03-2006 00:18	5	http://www.atlantasilverbacks.com
1038	moses sakyi	12-03-2006 00:21	10	http://www.afrique-sport.com
1038	liam george	12-03-2006 00:22	1	http://www.sacfc.co.uk

1038	liam george	12-03-2006 00:22	2	http://www.lutonfc.com
1038	liam george	12-03-2006 00:22	4	http://www.kickinmagazine.ie
1038	liam george	12-03-2006 00:22	9	http://www.pbase.com
1038	optygen	12-03-2006 12:56	1	http://www.prolithic.com
1038	when should i take optygen	12-03-2006 12:57	1	http://www.firstendurance.com
1038	pirlo	12-03-2006 15:29	1	http://www.andreapirlo.com
1038	pirlo	12-03-2006 15:29	7	http://www.uefa.com
1038	pablo maestroeni	12-03-2006 15:33	3	http://www.soccerconditioning.net
1038	castro ghana	12-03-2006 19:59	1	http://www.ghana.co.uk
1038	castro ghana	12-03-2006 19:59	2	http://www.ghana.co.uk
1038	castro ghana	12-03-2006 19:59	1	http://www.ghana.co.uk
1038	castro ghana	12-03-2006 19:59	9	http://www.ghanaweb.com
1038	makelele	12-03-2006 20:05	2	http://www.football-rumours.com
1038	makelele	12-03-2006 20:05	3	http://soccernet.espn.go.com
1038	joe afful	12-03-2006 20:06	3	http://digital-diva.smugmug.com
1038	joe afful	12-03-2006 20:06	10	http://www.atlantasilverbacks.com
1038	joe afful	12-03-2006 20:10	4	http://www.montrealimpact.com
1038	joe afful	12-03-2006 20:10	6	http://www.atlantasilverbacks.com
1038	ray goodlet	12-03-2006 23:13	1	http://www.sick-boy.com
1038	caleb norkus	12-03-2006 23:17	1	http://www.charlestonbattery.com
1038	caleb norkus	12-03-2006 23:17	6	http://www.matchnight.com
1038	caleb norkus	12-03-2006 23:19	10	http://www.cybersoccernews.com
1038	how to become a player agent	12-03-2006 23:44	4	http://www.thefa.com
1038	colo colo	12-03-2006 23:54	1	http://www.colocolo.cl
1038	colo colo	12-03-2006 23:54	2	http://www.csdcolocolo.com
1038	spanish translation	12-03-2006 23:57	1	http://www.freetranslation.com
1038	spanish translation	12-03-2006 23:57	6	http://www.trustedtranslations.com
1038	spanish translation	12-03-2006 23:57	10	http://www.jb-translator.com
1038	puerto rico islanders	13-03-2006 00:00	3	http://www.uslsoccer.com
1038	1999 hyundai accent	13-03-2006 00:03	9	http://www.internetautoguide.com
1038	credit suisse	13-03-2006 12:02	1	http://www.credit-suisse.com
1038	nba live 05	13-03-2006 21:40	5	http://cheats.ign.com
1038	nba live 05	13-03-2006 21:41	7	http://answers.yahoo.com
1038	nba live 05	13-03-2006 21:41	10	http://www.gamespot.com
1038	nba live 05	13-03-2006 21:41	14	http://www.cheatgenius.co.uk
1038	nbalive 2005 cheats	13-03-2006 21:52	3	http://www.cheatserver.com

1038	nbalive 2005 cheats	13-03-2006 21:52	5	http://www.gamespot.com
1038	nbalive 2005 cheats	13-03-2006 21:52	7	http://www.gamewinners.com
1038	nbalive 2005 cheats	13-03-2006 21:52	9	http://www.captaincode.com
1038	nbalive 2005 cheats	13-03-2006 21:52	10	http://www.cheatscodesguides.com
1038	nbalive 2005 cheats	13-03-2006 22:08	1	http://cheats.gamespy.com
1038	league.com	14-03-2006 02:25	3	http://www.uslfans.com
1038	www.adultsallowed.com	14-03-2006 14:33	1	http://www.adultsallowed.com
1038	www.hotmail.com'	14-03-2006 14:34	1	http://www.hotmail.com
1038	10q's	14-03-2006 20:29	4	http://www.lmlpayment.com
1038	10q's	14-03-2006 20:29	4	http://www.lmlpayment.com
1038	what are 10q's	14-03-2006 20:31	6	http://investor.delphi.com
1038	what are 10q's	15-03-2006 09:26	3	http://www.lmlpayment.com
1038	wall street journal	15-03-2006 09:30	9	http://www.collegejournal.com
1038	wall street journal	15-03-2006 09:33	1	http://www.msnbc.msn.com
1038	finance dictionary	15-03-2006 09:34	2	http://www.investopedia.com
1038	liam george	15-03-2006 16:25	10	http://www.pbase.com
1038	pedro power	15-03-2006 19:41	2	http://rockathletics.collegesports.com
1038	pedro power	15-03-2006 19:41	3	http://profile.myspace.com
1038	pedro power	15-03-2006 19:41	4	http://www.infosportinc.com
1038	diego walsh	15-03-2006 19:45	1	http://www.mlsnet.com
1038	how to dye dark hair blonde	15-03-2006 19:50	1	http://www.soyouwanna.com
1038	danny rawsthorne	16-03-2006 21:46	1	http://www.harchester.net
1038	gyrotonic transformer 1000	16-03-2006 23:46	1	http://www.asseenontv.com
1038	robbie savage	18-03-2006 13:12	8	http://en.wikipedia.org
1038	robbie savage	18-03-2006 13:12	10	http://www.football-rumours.com
1038	robbie savage	18-03-2006 13:14		
1038	robbie savage	18-03-2006 13:14	1	http://soccernet.espn.go.com
1038	martin jol	18-03-2006 13:16	1	http://www.4thegame.com
1038	martin jol	18-03-2006 13:16	10	http://www.martinjol.com
1038	lumiscope 2000-t	18-03-2006 13:22	1	http://shop.store.yahoo.com
1038	lumiscope 2000-t	18-03-2006 13:22	3	http://www.lumiscope.net
1038	lumiscope 2000-t	18-03-2006 13:22	5	http://www.southwestmedical.com
1038	lumiscope 2000-t	18-03-2006 13:22	7	http://www.rtamedicalsupply.com
1038	uses of lumiscope 2000-t	18-03-2006 13:25	1	http://www.liberatormedical.com
1038	joe afful	18-03-2006 13:26	3	http://www.atlantasilverbacks.com
1038	joe afful	18-03-2006 13:26	10	http://www.usldiscussions.com

1038	knee rehab	19-03-2006 11:36	1	http://www.kneeguru.co.uk
1038	knee rehab	19-03-2006 11:36	2	http://www.kneeguru.co.uk
1038	gavin glinton	19-03-2006 13:01	1	http://www.cnnsi.com
1038	gavin glinton	19-03-2006 13:01	3	http://www.matchnight.com
1038	gavin glinton	19-03-2006 13:01	10	http://soccer.azplayers.com
1038	gavin glinton	19-03-2006 13:01	9	http://und.collegesports.com
1038	ugo okoye	19-03-2006 13:07	10	http://www.charleston.net
1038	armando romero	19-03-2006 13:08		
1038	diadora	19-03-2006 13:20	3	http://www.diadoraamerica.com
1038	aisha buttons	20-03-2006 14:15	5	http://www.aishamusic.com
1038	georgia form 500	20-03-2006 14:21	5	http://www.taxengine.com
1038	georgia form 500	20-03-2006 14:21	6	http://www.conectared.com
1038	georgia form 500	20-03-2006 14:21	9	http://www.chiff.com
1038	georgia form 500	20-03-2006 14:24	2	http://www.etax.dor.ga.gov
1038	georgia form 500 2005	20-03-2006 14:25	2	http://www.etax.dor.ga.gov
1038	optygen	20-03-2006 14:50	1	http://www.prolithic.com
1038	optygen	20-03-2006 14:50	2	http://www.firstendurance.com
1038	sit in on discusions	20-03-2006 17:08	10	http://www.anetforums.com
1038	machel millwood	20-03-2006 17:44	1	http://soccernet.espn.go.com
1038	machel millwood	20-03-2006 17:44	3	http://www.baltimoreblast.com
1038	machel millwood	20-03-2006 17:44	5	http://www.rhinosfan.com
1038	machel millwood	20-03-2006 17:44	6	http://www.rhinosfan.com
1038	machel millwood	20-03-2006 17:44	7	http://umassathletics.collegesports.com
1038	machel millwood	20-03-2006 17:44	8	http://www.californiacougars.net:16080
1038	machel millwood	20-03-2006 17:48	10	http://www.rhinosfan.com
1038	jonathan steele	20-03-2006 17:49		
1038	jonathan steele	20-03-2006 17:49		
1038	lenin steenkamp	20-03-2006 17:51	5	http://www.soccersam.com
1038	jordan chirico	20-03-2006 17:53	1	http://iuhoosiers.collegesports.com
1038	jordan chirico	20-03-2006 17:53	3	http://www.fansonly.com
1038	ivo ilarionov	21-03-2006 11:33	3	http://www.atlantasilverbacks.com
1038	ivo ilarionov	21-03-2006 11:33	7	http://www.lynxsoccer.com
1038	ivo ilarionov	21-03-2006 11:34	5	http://www.ajc.com
1038	padraig drew	21-03-2006 11:36	8	http://www.dundalkfc.com
1038	padraig drew	21-03-2006 11:36	9	http://home.clara.net
1038	padraig drew	21-03-2006 11:37	3	http://www.kickinmagazine.ie

1038	padraig drew	21-03-2006 11:37	5	http://archives.tcm.ie
1038	padraig drew	21-03-2006 11:37	8	http://indigo.ie
1038	padraig drew	21-03-2006 11:37	11	http://www.fai.ie
1038	padraig drew	21-03-2006 11:39	11	http://graphics.fansonly.com
1038	hyundai accent 98 air bag	21-03-2006 14:33	7	http://cgi.ebay.com
1038	hyundai accent 98 parts	21-03-2006 14:35	3	http://www.autopartswarehouse.com
1038	airbags	21-03-2006 14:38	6	http://www.cars.com
1038	1999 hyundai accent air bag	21-03-2006 14:43	10	http://www.usedpartslive.com
1038	1999 hyundai accent air bag	21-03-2006 14:45	3	http://www.2carpros.com
1038	1999 hyundai accent air bag	21-03-2006 14:45	8	http://www.autobytel.com
1038	1999 hyundai accent air bag	21-03-2006 14:45	13	http://www.carsearch.com
1038	1999 hyundai accent air bag	21-03-2006 14:52		
1038	auto parts	21-03-2006 14:55	1	http://www.advanceautoparts.com
1038	auto parts	21-03-2006 14:58	3	http://www.autozone.com
1038	auto parts	21-03-2006 14:58	2	http://www.napaonline.com
1038	driver airbag	21-03-2006 15:01	1	http://www.autoliv.com
1038	driver airbag	21-03-2006 15:01	2	http://www.autoliv.com
1038	driver airbag	21-03-2006 15:01	4	http://www.cartoonstock.com
1038	driver airbag	21-03-2006 15:01	5	http://www.dft.gov.uk
1038	hyundai accent airbag pair air bags airbags	21-03-2006 15:11	1	http://www.qaparts.com
1038	kendall nkrumah	21-03-2006 15:19	8	http://scoreboards.aol.com
1038	afful	21-03-2006 15:20	14	http://www.fansonly.com
1038	hyundai dealer	21-03-2006 16:11	1	http://www.automotive.com
1038	hyundai dealer	21-03-2006 16:11	6	http://www.hyundaiusa.com
1038	alfredo esteves	21-03-2006 16:27		
1038	alfredo esteves	21-03-2006 16:27	15	http://www.ussoccer.com
1038	alfredo esteves	21-03-2006 16:29	10	http://www.socceramerica.com
1038	farfan	21-03-2006 22:27	4	http://www.farfan.tk
1038	farfan	21-03-2006 22:27	8	http://en.wikipedia.org
1038	eric addo	21-03-2006 22:28	2	http://en.wikipedia.org
1038	eric addo	21-03-2006 22:29	6	http://www.ghanaweb.com
1038	rans addo	21-03-2006 22:29	4	http://soccernet.espn.go.com

A fragment of Concept Instance File (PROBASE) containing adequate amount of entries is shown in Table C 1.1. The table has three columns. First and second column indicates concept, instance/ entity respectively and third column indicates the number of associations between concept and its corresponding instance.

Table C 1.1 Concept Instance File

Concept	Entity/ Instance	Number of association
factor	age	35167
free rich company datum	size	33222
free rich company datum	revenue	33185
state	california	18062
supplement	msm glucosamine sulfate	15942
factor	gender	14230
factor	temperature	13660
metal	copper	11142
issue	stress pain depression sickness	11110
variable	age	9375
information	name	9274
state	new york	8925
social medium	facebook	8919
material	plastic	8628
supplemental material	cds	8175
supplemental material	access code	8133
state	texas	8056
supplemental material	info trac	8006
detailed business	key executive	7979
detailed business	financials	7942
state	florida	7836
company	google	7816
material	metal	7809
parameter	temperature	7490
testing device	glucometer diabetes blood sugar test	7138
material	glass	6950
factor	size	6709
symptom	headache	6620
social medium	twitter	6589
condition	diabetes	6493
factor	stress	6433
metal	aluminum	6433
sport	basketball	6423
symptom	nausea	6364
heavy metal	lead	6361
fruit	apple	6315
factor	education	6256
city	new york	6251
sport	football	6244
symptom	fatigue	6206
environmental factor	temperature	6195
company	microsoft	6189
information	address	6136
natural disaster	earthquake	5894
place	parking place	5822
factor	cost	5661
side effect	nausea	5604
parameter	ph	5581
factor	smoking	5532

natural disaster	flood	5525
material	aluminum	5518
company	ibm	5494
inorganic contaminant	salt	5468
search engine	google	5430
food	fruit	5421
event	wedding	5393
complication	infection	5294
word	anticipate	5290
place	shop	5256
factor	location	5247
social medium site	facebook	5227
factor	diet	5205
event	sales event	5189
factor	ph	5160
personal information	name	5155
metal	iron	5124
chronic disease	diabetes	5111
inorganic contaminant	metal	5104
emotion	anger	5064
vegetable	carrot	5042
site	facebook	5008
social network	facebook	4987
symptom	fever	4966
sport	tennis	4964
fruit	strawberry	4824
animal	dog	4809
activity	fishing	4789
sport	soccer	4752
food	vegetable	4638
complex carbohydrate	whole wheat bread	4620
factor	weather	4604
variable	gender	4594
metal	nickel	4564
material	steel	4561
characteristic	age	4494
symptom	pain	4493
activity	swimming	4487
alcohol	ethanol	4466
information	date	4463
state	new jersey	4447
browser	firefox	4369
heavy metal	cadmium	4354
company	apple	4353
heavy metal	mercury	4326
metal	zinc	4313
descriptive statistic	mean	4271
social networking site	facebook	4270
mineral	calcium	4246
heat source	radiator	4227
state	illinois	4221
personal information	address	4218
market	china	4198
metal	lead	4184
factor	weight	4157
big deal	real estate investment opportunity	4135
risk factor	smoking	4134
warranty term	basic warranty	4131
warranty term	powertrain warranty	4131
warranty term	tires warranty	4131
warranty term	towing warranty	4131

alcohol	methanol	4092
metal	gold	4058
natural disaster	hurricane	4055
engine	google	4041
city	chicago	4033
polynucleotide sequence	est sequence	4017
metal	silver	3989
property type	house	3893
material	stainless steel	3887
property type	condo	3877
state	arizona	3846
factor	genetic	3844
material	paper	3828
fossil fuel	coal	3813
fruit	banana	3808
word	anticipates	3774
factor	climate	3756
food	meat	3756
information	location	3746
fish	salmon	3733
item	furniture	3730
factor	income	3727
item	clothing	3721
european country	germany	3717
food	fish	3711
factor	ethnicity	3708
demographic variable	age	3703
mobile device	smartphone	3700
vegetable	broccoli	3680
food	egg	3667
sport	baseball	3627
economy	china	3603
issue	climate change	3593
industry	construction	3580
food	potato	3576
state	ohio	3570
state	washington	3566
social medium platform	facebook	3556
microorganism	bacterium	3554
state	massachusetts	3548
liquid	water	3542
word	expect	3528
state	oregon	3519
organ	heart	3505
fruit	cherry	3485
occasion	wedding	3480
utility feature	brochure printing facility	3475
information	age	3465
medical condition	diabetes	3458
mineral	magnesium	3446
fixture	sink	3434
industry	automotive	3432
language	spanish	3431
vegetable	tomato	3390
european country	france	3386
social medium site	twitter	3357
dairy product	milk	3348
activity	hiking	3346
professional	doctor	3334
industry	pharmaceutical	3311
market	india	3305

language	french	3276
city	los angeles	3270
crop	corn	3243
item	cost of living	3219
state	colorado	3217
city	london	3210
animal	cat	3201
fixture	bathtub	3181
material	ceramic	3174
risk factor	age	3154
heavy metal	copper	3150
crop	wheat	3132
gas	nitrogen	3131
material	copper	3118
city	san francisco	3116
company	amazon	3115
additive	antioxidant	3084
material	concrete	3080
variable	temperature	3078
state	minnesota	3070
symptom	dizziness	3049
nutrient	nitrogen	3047
gas	carbon dioxide	3039
industry	manufacturing	3037
mineral	iron	3026
fruit	orange	3020
animal	rabbit	3012
organ	kidney	3011
state	virginia	3009
solvent	acetone	2998
animal	horse	2997
state	pennsylvania	2996
factor	obesity	2975
nation	china	2960
descriptive statistic	frequency	2957
sector	agriculture	2955
information	time	2953
area	education	2949
economy	india	2949
food	milk	2948
factor	humidity	2945
sport	golf	2944
factor	time	2940
inert gas	nitrogen	2937
condition	temperature	2928
factor	culture	2904
developed country	united states	2901
mobile device	tablet	2895
item	jewelry	2894
social medium platform	twitter	2894
food	bread	2890
occasion	birthday	2879
information	phone number	2861
term	organization listing	2860
professional	lawyer	2857
industry	mining	2848
item	book	2845
state	maryland	2839
inert gas	argon	2836
fossil fuel	oil	2811
food	cheese	2801

side effect	headache	2780
animal	bird	2779
event	school event	2775
credit card	visa	2771
dairy product	cheese	2746
environmental condition	temperature	2743
word	plan	2739
chronic disease	heart disease	2737
item	food	2716
industry	aerospace	2713
material	stone	2709
area	health	2703
social network	twitter	2698
metal	titanium	2690
factor	environment	2688
demographic datum	age	2682
word	estimate	2674
service	plumbing	2674
antioxidant	vitamin e	2674
issue	education	2668
technology	internet	2667
issue	security	2646
site	twitter	2646
outdoor activity	hiking	2643
chronic condition	diabetes	2630
organ	lung	2622
vegetable	cabbage	2613
microbial contaminant	virus	2596
sport	swimming	2591
company	facebook	2589
browser	chrome	2587
sport activity	tennis	2587
demographic characteristic	age	2579
risk factor	hypertension	2578
spice	cinnamon	2572
fruit	pear	2567
grain	wheat	2560
industry	banking	2548
characteristic	gender	2547
state	georgia	2545
vegetable	onion	2542
demographic factor	age	2541
appliance	refrigerator	2539
condition	arthritis	2537
grain	barley	2529
fluid	water	2506
utility	water	2502
asian country	china	2496
event	sports event	2495
metal	cadmium	2491
issue	health	2480
vegetable	spinach	2474
symptom	anxiety	2470
condition	asthma	2469
cruciferous vegetable	broccoli	2469
financial institution	bank	2465
issue	safety	2461
metal	stainless steel	2458
event	concert	2457
industry	healthcare	2455
city	boston	2454

crop	maize	2449
word	intend	2444
animal	sheep	2438
small portion	couple small cookie	2438
device	tablet	2436
event	school social event	2433
demographic information	age	2433
risk factor	obesity	2431
metal	chromium	2429
website	facebook	2428
nonsteroidal anti	ibuprofen	2427
activity	sport	2422
food	chocolate	2401
state	michigan	2394
information	email address	2393
search engine	yahoo	2387
activity	yoga	2386
issue	depression	2382
company	intel	2375
activity	cycling	2374
institution	school	2369
fruit	pineapple	2358
animal	deer	2357
symptom	depression	2354
fatty fish	salmon	2352
fruit	peach	2347
sport	volleyball	2339
nutrient	phosphorus	2326
industry	retail	2320
preciou metal	gold	2320
animal	cow	2318
material	rubber	2310
material	leather	2303
word	project	2301
food	bean	2301
side effect	dizziness	2298
mineral	zinc	2298
international organization	world bank	2291
sector	construction	2289
microsoft hardware failure	bad hard drive	2281
condition	depression	2280
fruit	grape	2277
metal	steel	2271
risk factor	diabetes	2270
place	brighton	2269
lifestyle factor	smoking	2269
factor	type	2268
industry	food	2266
personal information	phone number	2266
personal protective	glove	2260
mineral	potassium	2256
antioxidant	vitamin c	2255
activity	art	2247
skill	communication	2247
state	nevada	2246
nsaid	ibuprofen	2243
industry	chemical	2230
holiday	christmas	2217
sport	hockey	2212
stimulant	caffeine	2212
factor	experience	2211

web browser	internet explorer	2209
demographic variable	gender	2207
additive	plasticizer	2206
state	north carolina	2200
device	printer	2200
browser	internet explorer	2192
mechanical property	tensile strength	2190
outdoor activity	camping	2190
additive	pigment	2187
industry	textile	2185
gas	oxygen	2185
oily fish	salmon	2182
service	water	2181
descriptive statistic	percentage	2180
factor	lifestyle	2177
natural fiber	cotton	2172
food	pasta	2171
crop	rice	2171
relaxation technique	meditation	2166
pet	dog	2166
microbial contaminant	bacterium	2165
citrus fruit	orange	2164
profession	medicine	2157
company	dell	2154
subject	history	2154
vegetable	potato	2146
sector	education	2145
place	cornwall	2139
state	wisconsin	2136
animal	goat	2136
substance	alcohol	2131
autoimmune disease	rheumatoid arthritis	2128
outdoor activity	fishing	2126
activity	horse riding	2122
activity	kayaking	2121
negative emotion	anger	2121
material	sand	2109
dental service	filling	2109
grain	brown rice	2103
herb	basil	2090
renewable energy source	solar	2089
activity	canoeing	2080
service	physical	2080
activity	golf	2074
factor	quality	2065
natural disaster	tornado	2055
heavy metal	zinc	2044
fruit	mango	2039
activity	skiing	2034
medium	television	2033
factor	price	2032
item	toy	2032
loan option	maximum loan amount	2029
medication	antidepressant	2028
loan option	maximum amount borrowed	2026
nation	india	2021
electronic device	cell phone	2018
service	education	2015
skill	problem solving	2010
business	restaurant	2007
complication	bleeding	2005

solution	plumbing	2001
medication	aspirin	1998
crop	cotton	1998
subject	science	1997
personal information	social security number	1994
browser	google chrome	1993
industry	oil	1991
solvent	benzene	1991
utility	electricity	1990
food	rice	1986
amenity	restaurant	1983
additive	stabilizer	1980
variable	education	1979
herb	rosemary	1974
descriptive statistic	standard deviation	1974
small dog breed	toy	1974
beverage	coffee	1974
factor	health	1972
profession	law	1971
place	isle of white	1963
activity	reading	1952
material	brick	1951
institutional investor	pension fund	1942
pet	cat	1939
material	paint	1936
activity	tennis	1929
city	seattle	1929
metal	mercury	1928
factor	religion	1927
solvent	methanol	1923
activity	craft	1922
skin condition	eczema	1921
platform	facebook	1921
relaxation technique	yoga	1919
solvent	toluene	1915
problem	depression	1915

BRIEF PROFILE OF THE RESEARCH SCHOLAR

Mamta Kathuria has received her Masters in Computer Application from Kurukshetra University, Kurukshetra in the year 2005 and M. Tech. in Computer Engineering from Maharishi Dayanand University, Rohtak in the year 2008 respectively. She is pursuing her Ph.D in Computer Engineering from YMCA University of Science and Technology, Faridabad. She is currently working as an Assistant Professor in YMCA University of Science & Technology, Faridabad and has eleven years of experience. She has published more than twenty research papers in various international journals and conferences. Her areas of interest are search engines, Web Mining and Fuzzy Logic.

Mamta Kathuria
Assistant Professor
Department of Computer Engineering
Faculty of Engineering & Technology
YMCA University of Science and Technology
Sector 6, Faridabad (Haryana), INDIA
Tel: +91 9466369331
Fax: 0129 2242143
mamtakathuria31@gmail.com

LIST OF PUBLICATIONS OUT OF THESIS

List of Published Papers

Sl. No.	Title of paper	Name of Journal where published	No.	Volume & Issue	Year/ Page	Remarks
1.	Creation of Entity Synonyms Dictionary and its usage for Query Reformulation : A Review	Journal of Emerging Technologies and Innovative Research (JETIR)	(ISSN-2349-5162)	Volume 5, Issue 8	Aug 2018, 1185-1190	UGC (63975)
2.	A Fuzzy Logic based Synonym Resolution Approach for Automated Information Retrieval	International Journal of Semantic Web and Information System, IGIGlobal Publisher	DOI: 10.4018/IJSWIS.2018100105 210117-093922	IJSWIS: Volume 14, Issue 4, Article 5	Oct-Dec 2018,92-109	SCIE(Web of Science), SCOPUS, UGC(7790), ACM
3.	Discovery of Entity Synonym Using Anchor Text and URLs	International Journal of Future Generation Communication and Networking, SERSC publisher	http://dx.doi.org/ 10.14257/ ijfgcn.2017.10.11.03	Volume 10, No. 11	2017, 19-36	ESCI, UGC(22814), SCOPUS
4.	A Journey of Web Search Engines: Milestones, Challenges & Innovations	I.J. Information Technology and Computer Science, MECS Publisher	DOI: 10.5815/ ijitcs.2016.12.06	IJITCS Volume 8, No. 12	Dec. 2016, 47-58	Free Publication, Indexed in Stanford University Libraries, Cornell University Library
5.	Application of fuzzy logic in web mining domain: A survey	International journal of advance research in IT and Engineering, IJARIE, Greenfield advanced research publishing house, IJARIE	ISSN: 2278-6244	Volume 1, No. 3,	September 2012,| 1- 16	Google Scholar

List of Publications in International Conference

S_No	Title of the paper	Publisher	Impact factor	Whether Referred or Non Referred	Whether you paid any money or not for publication	Remarks
1.	Semantic Similarity between Terms for Query Suggestion	International Conference on Reliability, Infocom Technologies and Optimization(ICRITO),IEEE Conference	-	-	Yes	SCOPUS INDEXED
2.	A Survey of Semantic Similarity Measuring Techniques for Information Retrieval	International Conference on "Computing for Sustainable Global Development	-	-	Yes	SCOPUS INDEXED

List of communicated papers

S_No	Title of the paper	Publisher	Impact factor	Whether Referred or Non Referred	Whether you paid any money or not for publication	Present Status/ Remarks
1.	A Comprehensive approach to Dynamic Entity Resolution and Fuzzy Classification	Multimedia Systems	1.703	Referred	No	Under Review/ SCIE, Springer
2.	Concept Resolution for Focused and Enriched Web Information Retrieval	Interacting with Computers, Oxford University Press	1.410	Referred	No	Awaiting Reviewer Invitation after submitting 2nd revision/ SCI indexed

A FUZZY LOGIC BASED FRAMEWORK
FOR
RELEVANT INFORMATION RETRIEVAL

(SUMMARY)

1.1 GENERAL

The World Wide Web **(*WWW*)[1]** is a gigantic repository that keeps information related to almost every domain of knowledge accessible everywhere on anytime basis. The massive size, continuous update of the information, heterogeneity on the basis of various factors like linguistics, geographical location, cultures and other parameters make the task of information retrieval quite complex and challenging.

Though the performance of search engines in today's scenario is quite impressive yet there has been the ever felt need for novel mechanisms for accomplishing expectations of the users who are seeking the rich set of relevant results for their submitted queries.

The basic reasons for the inability of the search engines to provide the relevant results which are not up to expected levels are as follows:

- Query is a very short piece of text in natural language and successful retrieval is very much dependent of the intent of the user behind the query.
- Natural language is ambiguous and affects the relevance/quality of search results returned by the search engine.
- Users may use slang terms which are not as such part of the language.
- The reference in the query may be conceptual requiring proper instantiation.
- The reference in the query may refer to an entity recognizable by different names.
- Current Lexical resources are unable to cover the heterogeneity of the web.
- Web is continuously updating.

All these issues need to be addressed for getting the rich and relevant information from the web. The literature contains a lot of work in this regard, the study of which motivated us to carry out the work briefly described in this summary.

1.2 PROBLEM IDENTIFICATION

To understand the ongoing work being carried out to overcome the above mentioned problems, a lot of literature was were studied. It was felt that there is an ample opportunity to carry out further research to ensure the rich and relevant information by working on various components of the query. The main focus of the work done is to improve the query used by the user for getting relevant and useful results. It has been observed that queries to search engines on the Web are usually short, imprecise and ambiguous. They do not contain enough signals for statistical inference and do not provide satisfactory information for an effective selection of relevant documents. To find the relevant documents, matching is done purely on the basis of occurrence of keyword or expression in the document. But, it is not always necessary that a document containing a word with high frequency will be relevant. Thus, it can be seen that the relevance computation is done purely on the basis of occurrence of keywords in content; it does not consider the context of keywords. The proposed work in this thesis finds the equivalent word of query in the presence of contexts. So, the work paves the attention to identify the different categories of the query.

The literature survey has shown that the basic constituents of a query can be classified into four categories: *Keywords, Attributes*, *Entities and Concepts*.

(i) *Keywords* are non-trivial words which carry the essence of the query. The keywords make the query meaningful and are the major guiding factors for relevant information retrieval to be carried out by the search engines.

(ii) *Entities* are persons/places/objects referred in the query which have distinct and independent existence. Different users may refer to the same entity in different manners. For example, the newspaper *The Times of India* may also be referred to as *TOI*. A search engine must be able to handle these multiple versions of the references used in the query. These multiple references have been referred to as *entity synonyms* in the work [2, 3]. Creation of appropriate set of entity synonyms for a given entity is

also a major requirement for relevant and rich information retrieval. Various contributions in this field are available in [2-10].

(iii) *Attributes* are the words which define the features/ characteristics of entities and keywords used in the query. To enrich the search process, a web search engine may create multiple versions of the same query by using the appropriate set of synonyms of the attributes used in the query and create an index to access the quality of the synonym generated. Various contributions in this field are available in [11-19].

(iv) A *Concept* in the query is a word which refers to a broad category of objects in generic manner. For example, in the query "*good actors in India", Good actors* is a concept. A concept referred in query has to be translated to its closest set of instance(s). Handling of the concept is the most challenging task for the search engine as its resolution requires the understanding and usage of worldly knowledge. The instantiation of a concept can vary depending upon the local & global contexts. The hardest part of the query expansion is to find the appropriate instantiation for the concept used in the query as per the requirement of the user. Various contributions in this field are available in [20-31].

After going through the literature, following inferences were drawn:

- Keywords are the essential part of the query and should not be disturbed/modified/altered.
- Lexical resources are unable to provide the requisite set of synonyms for the words used in the query owing to the widespread and heterogeneous nature of the web. So, there is a need to find out global mechanisms for creating the set of synonyms which truly cover the heterogeneity of the web.
- Alternative references to entities (also known as entity synonyms) are not at all supported by the lexical resources. The only way one can find out the entity synonyms is through web exploration and analysis of web logs.
- Conceptual references need to be translated into their worthy instances which are quite a challenging task as it requires worldly knowledge.

After exploring all this literature, we were in a position to set the objectives of the proposed work.

1.3 OBJECTIVES

As the amount of information on the web is increasing and changing rapidly without any control, existing search methodologies do not fit best to find the required information. Therefore, the need arises to devise some new methodologies so that relevant data can be retrieved.

In the light of the above motivational factors, the main objective of the proposed work is to improve the performance of search engines using query recommendation so that relevant information can be retrieved from WWW.

Following objectives were set for the proposed work:

a) To devise a mechanism to search synonyms of an attribute word of the query using huge document repositories.
b) To devise a mechanism to search rich set of entity synonyms for an entity using static and dynamic web.
c) To design an index to assess the quality of synonyms as two synonyms of the same word can't have same intensity.
d) To devise a mechanism to translate a concept to its intended instances using worldly knowledge source.
e) To devise a mechanism for automated usage of identified set of synonyms to be utilized by the machine.

1.4 PROBLEM DEFINITION

To ensure relevant web search through query rephrasing or expansion using:

- rich set of identified synonyms for the entities and the attributes used in the query
- appropriate instances for the concepts present in the query

and to devise novel mechanisms for the purpose.

1.5 ACCOMPLISHMENTS

Following accomplishments were made during this work:

a) A mechanism to search synonyms of an attribute word on the basis of the context identification using multiple corpora was proposed and implemented. The method is quite an improvement over the existing methods based on page count and snippets. The proposed work is used in web search and other applications such as:
   i) Enrichment of lexical Resources
   ii) Word Sense Disambiguation
   iii) Automated query expansion for web search

b) A mechanism to generate rich set of entity synonyms for an entity using query log and anchor text was proposed and implemented. The technique is scalable and can be implemented for both unstructured and dynamic web. The work can be used for automated search process by the search engine using the techniques like Fuzzy Rule Base and Knowledge Graph etc. The basic outcomes are:
   i) Query expansion for enriched search without losing the relevance
   ii) Improved Search Relevance
   iii) Query Auto Suggestion
   iv) Creation of Entity Knowledge Graph

c) For both of the above mechanisms, an index was created to assess the quality of synonyms. The index was fuzzified and a Fuzzy Rule Base (FRB) was created for automated deployment of synonyms for various purposes. The proposed system enables to make the user to choose soft decision based on the extent of semantic similarity by the use of fuzzy logic rather than rejection in case of hard decision. The outcomes achieved are:
   i) Fuzzy Rule Base creation
   ii) Use of Fuzzy Rule Base for the purpose of automation by choosing the semantically similar word based upon the similarity index

d) A mechanism to translate a concept to its intended instances was proposed and implemented using PROBASE (the largest available worldly knowledge source) [32]. The approach can be considered as a mechanism from large

scale generalization created by the voluminous worldly knowledge to the specific requirement of the user. Outcomes achieved are:

(a) Supporting the machine intelligence with real world knowledge.

(b) Enrich the web search by moving from conceptualization to instantiation.

1.6 CONTRIBUTION

The proposed work contributes towards query expansion and resolution in existing search engines. It has been observed that for finding the relevant documents in response to user query, the traditional methods used by search engines are not a correct choice because they return documents in response to query keywords given by the user without considering the semantic aspect of a query.

In this work, specifically, a novel architecture for resolving different types of queries has been proposed and implemented. The proposed design is shown in Fig. 1. It takes the query as an input from the users and returns the recommendations for the proper replacement of the original input query.

The major functional components of the framework are given below:

a) Procedure for finding semantically similar word

The main objective of the work is to propose a synonym resolution method based upon the immediate context in various corpora. The search engines, if do not take into account the context of a query and unseeingly use the online lexical resources, may lead to fetching of large number of undesired pages which are otherwise not essential. To defeat this weakness, it is required that query be expanded through meaningful semantic expressions using context set. Keeping this need in view, this module focuses on a novel approach which tunes the lexical resources wherein the derived synonyms of a word are also provided with their applicable context. The outcome of the work includes context set identification for a word and computation of an index indicating the extent of semantic similarity between a pair of words.

The computed index has been fuzzified into a Fuzzy Rule Base for the purpose of automation and its usage into the web search engines and other such applications.

To ensure the unbiasedness of the approach, multiple corpora with wide range of genres and with each one containing huge set of documents have been taken up. The similarity index has been computed on the basis of commonality of the contexts. Various benchmark indices have been used to find the similarity index and the results have been normalized. The results obtained have been compared with a standard toolkit (UMBC) for the purpose of authentication. The obtained results have been used for the enrichment of lexical resources.

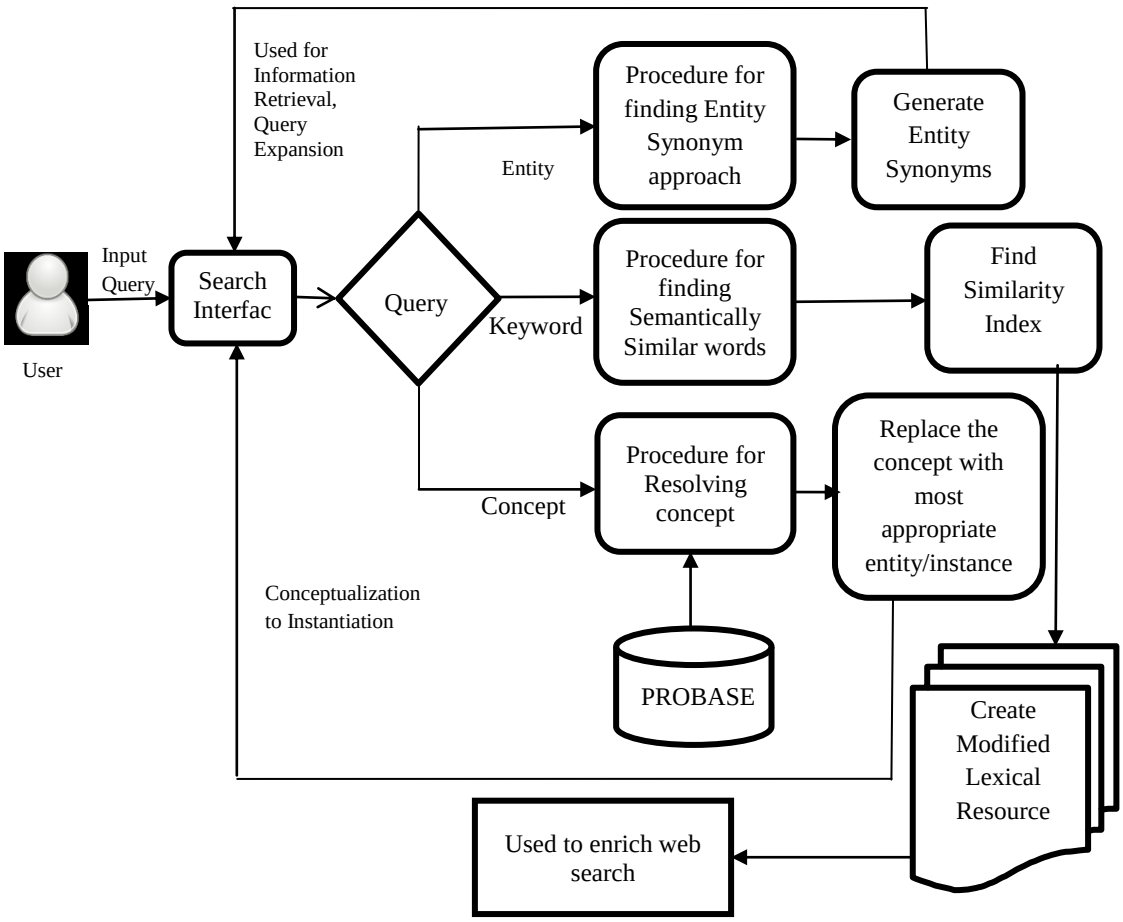


Fig 1: A Framework for Relevant Information Retrieval

b) Procedure for finding Entity Synonym

The proposed work involves the iterative use of Search Engine Result Pages (SERPs), extraction of context from the URL, query logs and extraction of anchor text. The process begins with the issuance of query by the web user on any search engine interface. The search engine receives the query and returns result pages referred to as

SERPs. In the proposed approach, the SERPs are searched in the query log to obtain the candidate synonyms. The title and snippets of the URLs of these SERPs are used to obtain the contexts. A new query is then issued to the search interface using a combination of a candidate synonym and the context related to an entity to obtain a new set of SERPs. An algorithm based on *Inbound Anchor Text* is then applied to extract more and more possible synonyms. The URLs obtained are collected to form parent URLs. Next, these parents URLs are further treated as input to generate sub parent URLs (SPUs). Then, each page of SPU is visited and all the (anchor text, link) pairs are collected in a hash map as a set of child URLs. Finally, each of these child URLs are compared with the parent URL and when exact match gets found, the corresponding anchor text is saved in another map. The title of the SPU and snippets are also used to explore more candidate synonyms.

After getting the candidate synonyms, similarity index is computed between the actual entity word and the candidate synonyms using *Web Jaccard* method. The index values obtained are then normalized between the range [0, 1]. Taking the normalized fuzzy value as the outline criteria, fuzzy sets are defined to express the quality of synonyms linguistically. These fuzzy sets are then used in Fuzzy Rule Base (FRB) for the automated application of entity synonyms in the web search process.

Entity synonyms generated through the proposed method have an edge over prevailing mechanisms, as it provides:

- More relevant set of entity synonyms (both in terms of quantity and quality)
- An index to access the quality of generated entity synonyms.
- Fuzzification of the Index for the purpose of automation.

The work will contribute to web-search in following ways:

- Improved search relevance
- Improved user experience
- Query auto suggestion
- Creation of entity dictionary
- Meaningful query expansion for the queries involving entities.

c) Procedure for Concept Resolution

The outlined framework consists of a set of modules for carrying out various functions. The very first module isolates the entities and concepts present in the input query by utilizing Concept Entity Relationship File (CERF). The CERF is created by referring PROBASE wherein each concept present in the input query is looked for the entities corresponding to the concept are picked up. CERF is populated by isolating concept from entities using tab and all entities related to concepts using comma operator. A concept list is thus generated by taking the substrings of concept and synonym of the concept present in the input query.

The concept list so obtained is then refined using concept synonym identification and their merger, by tracking IP address, using browsing history, query restructuring and through the use of typical associations. It checks the browsing history to locate the matching concepts, which acts as a source of query suggestion on the search interface for the user to help him/her in rephrasing the query in order to get the focused and intended results. The module then tracks the IP address of the user (based on geographic location of the user) to produce sub(sub)concept list.

The motivation behind the IP address usage is to go for the localized orientation of the query because normally one looks for the information identified with his/her local domain. For example, an actor for a U.S citizen is likely to be *hollywood actor* and for a indian citizen is likely to be a *bollywood actor*. For the concept list generated so far the sub-concepts are investigated which can in turn produce new sub-concepts or the entities. The entities generated are put into search engine result pages and the sub-concept investigation continues till they are at last changed over into their associated entities. This completes the concept resolution process.

After the concept resolution process, backtracking is used to distinguished entities which belong to majority of the concepts generated in the previous modules. An entity belonging to or having relationship with larger number of concepts is a probable candidate for instantiation. For this purpose ranking given in the *PROBASE* which gives the number of associations between the concept and instance on the web has likewise been utilized. The use of backtracking and number of associations can help in resolving the concept to their intended instances.

The algorithm offers a mechanism which provides large scale generalization created by the voluminous worldly knowledge to the specific requirement of the user. The proposed system is simple and is able to provide ease to web users to build a proper search query with the knowledge domain terminology which will help search engine to get the desired results.

1.7 CONCLUSION

As the volume of information on the WWW continues to increase on daily basis, almost all the information available in almost all the domains is accessible on it in today's scenario. Not only the volume of information is increasing on continuous basis, but more and more heterogeneity is also becoming the part of it as the contributions are coming from around the world involving linguistic, cultural and geographical differences.

The low cost of data usage and anywhere/ anytime availability of information has proved to be a motivating factor for seeking the information from the web instead of other resources like encyclopedia and libraries. But, the crux of the situation is:

Query is still a very short piece of text.

Exploring such a gigantic volume of information with the query text is becoming more and more challenging task for the search engines. So, there is a need to create different semantically similar versions of the query to cover the entire spectrum of the information sought. This can be done by finding out the semantically similar versions of the words used in the query (word synonyms) and similar names of entities (entity synonyms) which are compatible worldwide and have the capability to deal with the heterogeneity of the web. Also a concept used in the query must be translated to its appropriate set of instances in the light of worldly information. The work carried out in this thesis is an effort in this direction.

A summarization of the carried out work is as follows:

- The proposed technique is used to overcome the problems of synonymy and polysemy in the information retrieval field, by finding the semantically similar words with respect to the input query.

- The fuzzy sets created in the proposed work can be used for intelligent decision making by creating a Fuzzy Rule Base (FRB) that can be run on an appropriate fuzzy inference engine.
- It proposes and implements a credible method to generate a rich set of global entity synonyms for the commonly used entities using web data wherein the availability of the candidate data is not a priori requirement.
- Entity synonym finding technique is scalable and can be implemented for both unstructured and dynamic web.
- In information retrieval by suggesting the alternative name of the entity for getting more relevant set of documents in response to the user query.
- It helps to create Entity Dictionary or Entity Knowledge Graph that can be used to enhance search.
- It enables search engines to associate the concept used in query with appropriate set of instances using a worldly knowledge source called PROBASE in the light of the factors like user's browsing history, geographical location and IP address etc.
- It also discusses the textual practices for phrase sense disambiguation for meaningful web search.
- The proposed work deals with poor quality queries by finding most appropriate replacement of original query. Thus, it helps to discover relevant information as per user query.

It is hopeful that the proposed work shall be immense help to the information and computer science professionals.

1.8 FUTURE ENHANCEMENT

The work can be further extended by devising more refined methods which are able to take up the heterogeneity of the web in the simplistic and convincing manner and construct effective rephrased queries which cover the larger spectrum of the information. Also, more and more sources of worldly knowledge sources can be created and utilized to ensure the more effective translation of the concept to its appropriate set of instances.

Some of the possible extensions and issues that could be further explored in the near future are as follows:

- The automated search system based upon semantic similarity proposed in this dissertation can be extended to serve complex user queries, besides serving topical and informational queries.
- The work can be extended by the inclusion of more parameters and application of sophisticated techniques.
- The proposed method works on query elements. The query segmentation process has been left for the search engine and can be worked on in future.

1.9 ORGANIZATION OF THE THESIS

The thesis has been organized as follows:

Chapter II: Literature Survey: This chapter contains a discussion on the available work related to search engine evolution, semantic similarity between words, entities and concept based web search. Based on the literature survey on each topic, the problems and challenges have been identified and discussed in brief. These problems and challenges form the basis for the work carried out.

Chapter III: Synonym Resolution for Attribute Component in Query: This chapter talks about the proposed semantic similarity technique and its implementation for attributes component present in the query. To assess extent of similarity between the synonyms under consideration and the candidate word, list of their contexts have been taken into consideration. The work makes use of various corpora for extracting contexts.

Chapter IV: Dynamic Entity Resolution: This chapter covers the detailed discussion on the proposed and implemented work to generate a rich set of entity synonyms for the commonly used entities using web data, web log and anchor text. An index has also been created to assess the quality of the created synonym. Obtained results have been compared with the existing techniques.

Chapter V: Concept Resolution for Focused and Enriched Web Information Retrieval: This chapter proposes and implements an algorithm for concept resolution

using *PROBASE*, a huge taxonomy on worldly knowledge created by Microsoft, in combination with users' statistics resulting in focused and enriched outcomes. The results so obtained have been compared with the outputs of existing search engines such as Google, Bing and Yahoo.

Chapter VI: Conclusion and Future Scope: This chapter concludes the work and provides a description of potential future work in the area under consideration.

REFERENCES

[1] https://www.techopedia.com/definition/5217/world-wide-web-www

[2] Tao Cheng, Hady W. Lauw and Stelios Paparizos, "Entity Synonyms for Structured Web Search", *IEEE Transactions on Knowledge and data engineering*, Vol. 24, No. 10, pp. 1862 – 1875, October 2011.

[3] Hamid Mousavi, Shi Gao and Carlo Zaniolo, "Discovering Attribute and Entity Synonyms for Knowledge Integration and Semantic Web Search", *In Proceedings of the 3rd International Workshop on Semantic Search Over the Web*, Riva del Garda, Italy, ISBN: 978-1-4503-2483-0, August, 2013.

[4] Surajit Chaudhuri, Venkatesh Ganti and Dong Xin, "Mining Document Collections to Facilitate Accurate Approximate Entity Matching", *In PVLDB*, Journal Proceeding of VLDB Endowment, Vol 2, Issue 1, pp. 395-406, August 2009.

[5] Kaushik Chakrabarti, Surajit Chaudhuri, Tao Cheng and Dong Xin, "A Framework for Robust Discovery of Entity Synonyms", *In KDD '12 Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, Beijing, China, ISBN: 978-1-4503-1462-6, Pages 1384-1392 August 2012.

[6] Benjelloun, Omar, Hector Garcia-Molina, David Menestrina, Qi Su, Steven Euijong Whang, and Jennifer Widom, "Swoosh: A Generic Approach to Entity Resolution," *The VLDB J.*, Vol. 18, pp. 255-276, 2009.

[7] Bhattacharya, Indrajit, and Lise Getoor, "Collective Entity Resolution in Relational Data," *ACM Trans. Knowledge Discovery from Data(TKDD)* , Vol, 1 Issue 1, March 2007.

[8] L Jiang, Lili, Ping Luo, Jianyong Wang, Yuhong Xiong, Bingduan Lin, Min Wang, and Ning An, "An entity-relation graph based framework for discovering entity aliases" *In ICDM, IEEE 13th International Conference*, pp. 310-319, 2013.

[9] Yanen Li, Bo-June (Paul) Hsu, Cheng Xiang Zhai and Kuansan Wang, " Mining Entity Attribute Synonyms via Compact Clustering". *In CIKM '13 Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, San Francisco, California, USA, pp. 311-316, October 27 - November 01, 2013

[10] Roi Blanco, B. Barla Cambazoglu1, Peter Mika, and Nicolas Torzec, "Entity Recommendations in Web Search", *In International Semantic Web Conference*, ISBN: 978-3-642-41337-7, pp. 33-48, October 2013.

[11] Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka, "A Web Search Engine-Based Approach to Measure Semantic Similarity between Words", *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, NO. 7, pp. 977 – 990, July 2011

[12] Li, Yuhua, Zuhair A. Bandar, and David McLean, "An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources", *IEEE Trans. Knowledge and Data Eng.*, Vol. 15, No. 4, pp. 871-882, July-Aug. 2003.

[13] Adhikesavan, Kavitha, "An Integrated Approach for Measuring Semantic Similarity between Words and Sentences using Web Search Engine", *International Arab Journal of Information Technology, IAJIT*, Vol 12, issue 6, 2015.

[14] Bjerva, Johannes, and Robert Östling, "ResSim at SemEval-2017 Task 1: Multilingual word representations for semantic textual similarity", *In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 154-158, 2017.

[15] Elekes, Ábel, Martin Schäler, and Klemens Böhm, "On the Various Semantics of Similarity in Word Embedding Models", KIT – In Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries, IEEE Press, pp. 139-148, 2017

[16] Huda, Kohei Hayashi, and Danushka Bollegala, "An Optimality Proof for the PairDiff operator for Representing Relations between Words", *arXiv preprint arXiv*:1709.06673, 2017.

[17] Ranjbar, Niloofar, Fatemeh Mashhadirajab, and Mehrnoush Shamsfard, "Mahtab at SemEval-2017 Task 2: Combination of Corpus-based and Knowledge-based Methods to Measure Semantic Word Similarity", *In Proceedings of the 11th International Workshop on Semantic Evaluation*, (SemEval-2017), pp. 256-260, 2017.

[18] Recski, Gábor, Eszter Iklódi, Katalin Pajkossy, and Andras Kornai, A., "Measuring semantic similarity of words using concept networks",

In Proceedings of the 1st Workshop on Representation Learning for NLP, pp. 193-200. 2016.

[19] Alvarez, Marco A., and Seung Jin Lim. "A graph modeling of semantic similarity between words." In *null*, pp. 355-362. IEEE, 2007.

[20] Boucenna, Fateh, Omar Nouali, and Samir Kechid, "Concept-based Semantic Search over Encrypted Cloud Data", *In Proceedings of the 12th International Conference on Web Information Systems and Technologies*. Rome, Italy, pp. 235-242, 2016.

[21] Egozi, Ofer, Shaul Markovitch, and Evgeniy Gabrilovich, "Concept-Based Information Retrieval Using Explicit Semantic Analysis", *ACM Transactions on Information Systems*. Vol 29 Issue 2, Article No. 8. New York, NY, USA. pp. 1–34, April 2011.

[22] Fonseca, B.M., Golgher, P.B., Pôssas, B., Ribeiro-Neto, B.A., and Ziviani, N., "Concept-based query expansion", *In: CIKM*, 696–703, 2005.

[23] Lu, Yi-Jie, Phuong Anh Nguyen, Hao Zhang, and Chong-Wah Ngo, "Concept-Based Interactive Search System", *International Conference on Multimedia Modeling MMM*, Springer, Cham, pp. 463-468, 2017.

[24] Metzler, Donald, and W. Bruce Croft. "Latent concept expansion using markov random fields." *In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 311-318. ACM, 2007.

[25] Sendhil kumar, S., and T. V. Geetha., "Concept based Personalized Web Search", *TMRF, Advances in Semantic Computing*, Vol. 2. pp. 79- 102, 2010.

[26] Song, Yangqiu, Haixun Wang, Zhongyuan Wang, Hongsong Li, and Weizhu Chen, "Short text conceptualization using a probabilistic knowledgebase", *In: IJCAI,* , Vol 3, pp. 2330-2336, ISBN: 978-1-57735-515-1, 2011.

[27] Wang, Yue, Hongsong Li, Haixun Wang, and Kenny Q. Zhu, "Concept-Based Web Search", *International Conference on Conceptual Modeling*, Springer, Berlin, pp. 449-462, 2012.

[28] Wang, Zhongyuan, Kejun Zhao, Haixun Wang, Xiaofeng Meng, and Ji-Rong Wen, "Query Understanding through Knowledge-Based Conceptualization", Proceeding *IJCAI Microsoft Research*, pp. 3264-3270, July 2015.

[29] Wang, Fang, Zhongyuan Wang, Zhoujun Li, and Ji-Rong Wen. "Concept-based short text classification and ranking." *In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pp. 1069-1078. ACM, 2014.

[30] Zhang, Lanbo, "Interactive Retrieval Based on Wikipedia Concepts", *Arxiv Preprint arXiv*. 1412.8281, 2014.

[31] Recski, Gábor, Eszter Iklódi, Katalin Pajkossy, and Andras Kornai. "Measuring semantic similarity of words using concept networks." *In Proceedings of the 1st Workshop on Representation Learning for NLP*, pp. 193-200. 2016.

[32] Wang, Z., Huang, J., Li, H., Liu, B., Shao, B., Wang, H., Wang, J., Wang, Y., Wu, W., Xiao, J., Kenny Q. Z., "Probase: a Universal Knowledge Base for Semantic Search", *Microsoft Research Area*, May 2011.