International Journal of Information Retrieval Research
Volume 12 • Issue 3

A Hybrid Approach to Discover Entity Synonyms

Mamta Kathuria, J. C. Bose University of Science and Technology, India*
Chander Kumar Nagpal, J. C. Bose University of Science and Technology, India
Neelam Duhan, J. C. Bose University of Science and Technology, India

ABSTRACT

A major part of the current web information retrieval involves the fulfilment of people's daily needs based upon finding entities involving movies, events, persons, places, concepts, etc. The trend has been the outcome of the tremendous amount of precise and detailed information available on the web making it possible to retrieve the precisely targeted information involving specific instance of a particular domain called entity. The major problem in the entity-based search is the diversity of the name references made to the same instance of the same domain by the content creators and the searchers. These diversified references are known as entity synonyms. The problem is that they cannot be handled through lexical resources. This paper takes into account the previous works carried out in this direction and proposes an advanced mechanism that creates a richer set of entity synonyms than the previous approaches. The work also proposes an index to assess the quality of entity synonyms so generated which is further normalized and fuzzified for implementing automated search.

KEYWORDS

Entity Synonym, Entity Synonym Mining, Fuzzy Matching, Query Log, Synonym Discovery, Web Query

1. INTRODUCTION

With time the web is getting richer and richer with the enormous amount of data being uploaded on daily basis. The different types of users have been inputting the diverse types of queries on this data for getting the desired pin pointed information related to specific instance of a particular domain. For example, to search an information in today's scenario *The Times of India* newspaper there can be multiple types of queries such as "TOI today", "The Times of India today", "Today's TOI", "TOI 27/5", "TOI 27 May 2017" etc. Similar trends also appear in other real-world scenarios involving a variety of domains like stock markets, books, movies, cities etc. wherein the same entity is referred with different names. To mention a few, the movie *Dilwale Dulhania Le Jayenge* is more commonly known as *DDLJ*. The newspaper *The Hindustan Times* is more commonly known as *HT*. *Maruti*

DOI: 10.4018/IJIRR.300296 *Corresponding Author

800 is commonly known as *800*, *Sony Alpha Digital SLR* camera is commonly known as *Sony SLR*, Microsoft Excel is commonly referred to as MS-XL or simply XL.

An entity may be referred in multiple ways owing to differences caused by a variety of user groups, variations of spellings across regions or cultures, usage of abbreviations, typographical errors and other such reasons associated with conventional usage. Identification of a particular entity in a piece of text in the context of noisy data of the web is an uphill task for which no single methodology can be proposed as such. Therefore, the extraction and classification of the Entity Synonyms is an important area of research in the field of web information retrieval. It can help the search engine in understanding the text as a whole and map it to the appropriate real-world context. Furthermore, it will also create or determine the relationship between various real-world entities.

To enhance the user search experience it is required that all these semantically same queries referring to the same entity should be properly answered by targeting that particular entity. This problem can't be solved using dictionaries as there is a gap between the description of the end user and the way actual data has been described by the content creator. Lexicon based approaches like Stark & Riesenfeld (1998) focus on language-based alternatives that are not that applicable towards domains such as movies, company names, products, events, etc.

The web, due to its huge size of information repository, should be in a position to answer different types of queries submitted by the users in diverse forms. But the problem is more critical when the search is made over a specific entity as the content creator can't enter or describe all the known alternative names/ forms of that entity in the database. Thus to enhance the user search experience, there is a need to devise a mechanism to discover a list of entity synonyms for the common types of entities covered under the domain of *Entity Resolution.* The purpose of the domain is to reduce the gap between web users and content creator in the area of web search. The major problem with this domain is that the entity synonyms have a wide heterogeneous variety without following a particular trend making it quite a challenging task to discover all the entity synonyms for a specific entity.

Few possible mechanisms/techniques used for entity resolution include: techniques like acronyms and abbreviations, sources like Christian & Kjetil (2010) and Pei et.al.(2020) empirical methods based upon the web search and web log.

A work in the direction of Entity resolution shall lead to: improved search relevance, improved user experience, query auto-suggestion, creation of entity dictionary, and meaningful query expansion for the queries involving entities.

This paper is an effort in this direction and extends/enhances the work carried out by the earlier researchers in this domain.

2. ORGANIZATION OF THE PAPER

The paper is organized as follows: Section 3 consists of a survey on existing web-based empirical methods available in the literature to find out the entity synonym set for a given input string. Section 4 contains the definition of the problem and objectives set for the work. Section 5 shows the overview of the Experimental Setup. Section 6 contains details of the proposed methodology for entity resolution that provides better and enhanced results compared to its predecessors. Section 7 presents the results using table and graphs. Section 8 focuses on the hallmark of the proposed scheme. Section 9 discusses the usage of results. Section 10 talks about the conclusion and future scope of the work.

3. LITERATURE SURVEY

Entity synonyms are important ingredients of current web search as they improve search relevance and user experience making *entity resolution* a vital area of research. To gather entity synonyms, manually created knowledge bases like Freebase and Wikipedia can be used. Freebase consists of semantic knowledge and name aliases for most prominent entities. To collect valid entity synonyms

from Wikipedia, redirection pages and disambiguation pages are used. Strube & Ponzetto (2006) have talked about retrieving 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 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 being made to find web-based empirical methods to generate the entity synonyms. We take a look at some of these efforts and their limitations.

Hu et al.(2008) have 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 account the page content as a whole.

Chaudhuri et al.(2009) 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.

Malekian et al.(2008) 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 queries 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*.

Some papers Dey et al. (2002), Dong et al. (2005), Benjelloun et al. (2009), and Bhattacharya & Getoor (2007) have tried to find out the entity synonyms using the reconciliation process in the databases. They have used the different references of the same real-world entity in separate or similar databases as entity synonyms.

The major problem with the above citations is that they are unable to take into account the massive & heterogeneous content of the web. Moreover, in most 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. One such work has been published by Cheng et al. (2012).

Cheng et al. (2012) have 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*. 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 clicks on page *p* after issuing query *q*. Set *A* finds out the relevant 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 measures the strength of the relationship between the two candidate strings based upon their surrogates pages identified and 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 log 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.

Chakrabarti et al. (2012) have proposed a method to overcome the problems of click similarity Cheng et al. (2012) and document similarity Chaudhuri et al. (2009). 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 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.

Srikantiah et al. (2013) have proposed a mechanism to find the synonyms from the web based on 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 paper.

Xiang et al. (2015) proposed a new method of finding the entity synonyms. 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.

Mamta et.al.(2018) in their work intend to create a dictionary that can link different entity synonyms together. The work creates a sample database of entity synonyms and helps in reformulating the queries enabling the inclusion of documents where the entity being searched is having the alternate name.

Shen et al. (2019) proposed a new algorithm SynSetMine, that uses remote supervision from knowledge bases to build a set-instance classifier. The effectiveness and efficiency of SynSetMine has been demonstrated on three different benchmark datasets Wiki, NYT and PubMed.

Chenwei Zhang et al.(2020) proposed a neural network-based model named SYNONYMNET to check whether two entities are synonyms of each other or not by using different contexts. The experimental results show an improvement in mean average precision when compared with existing baseline models.

Pei et.al.(2020) proposed a novel set-aware Entity Synonym Discovery(ESD) which makes entities synonym sets effective in enabling the flexible receptive field for ESD. They have shown the efficacy of the proposed approach on public datasets with reasonable success.

Shen et.al.(2020) developed set expansion model to discover infrequent synonyms for popular entities resulting in increased recall of set expansion models. The results achieve comparable performances with the current state-of-the-art methods on existing Wiki and APR datasets.

Yang et.al. (2021) proposed a novel knowledge graph generator called KGSynNet to facilitate entity suggestions. A specifically designed fusion gate is utilized to adaptively incorporate the knowledge associated with entities into their semantic features to create a comprehensive representation of entities. They have claimed that their model outperforms the BERT(Bidirectional Encoder Representations from Transformers) model by 8.3% in terms of positive feedback.

The work discussed above suffers from following drawbacks.

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 references 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 account the page content as a whole.
6. There is no method for defining an index to assess the quality of synonyms generated.
7. The input mentions may be out-of-vocabulary and may come from a different semantic space of the entities.
8. Some entities rarely appear due to the long-tail effect.
9. Most of the approaches fail to take up the synonyms for general purpose common entities.

The above limitations identified in available work forms the basis of the proposed work.

4. PROBLEM IDENTIFICATION, DEFINITION AND OBJECTIVES

4.1 Problem Identification

There is a need to devise a mechanism that creates a rich and global set of entity synonyms based upon heterogeneous content of the web without the requirement of the candidate data. Identified synonyms should be based upon the context and an index be generated to assess the quality of synonyms.

4.2 Problem Statement

To devise 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 and to make a quantitative assessment for matching the similarity.

4.3 The Objectives

The objectives to be achieved are:

- Synonym set should be rich and global
- Automated generation of candidate synonyms
- Selection of synonyms should be context-based
- An index to assess the quality of synonyms generated
- A mechanism for automated application of entity synonyms for various purpose.
- Fuzzification of the Index for automation.

5. THE PROPOSAL

The proposed work comprises the iterative utilization of Search Engine Result Pages (SERPs), extraction of context from the URL, extraction of anchor text and candidate synonyms from query log. The proposed approach generates the set of entity synonyms using static and dynamic data. For static data, web query log is used as an offline source. For dynamic data, online web content is used. 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, the first level of candidate entity synonyms is obtained. Now, these initial sets of candidate synonyms are combined with contexts in 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 synonym and the context related to an entity to obtain a new set of SERPs.

For extracting rich and more focused entity synonyms, dynamic web content is used. For this purpose, the algorithm based on Inbound Anchor Text is applied. The anchor text is the clickable text

in a hyperlink and is relevant to the page a user is looking for, rather than generic text. The process begins whenever a new query that 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. Next, these parent 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 is compared with the parent URL and when a correct match is 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.

Thus, entity synonym is extracted using four things:

- From the user history log database
- From child map in case of match
- From trailing part of sub parent URL
- From combination of query and context obtained from child documents

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 deðned 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.

Figure 1 shows the basic methodology of the proposed work. Figure 2, Figure 3 & Figure 4 shows the architecture for different components of the proposed model.

(i) Candidate Synonym Extractor Module
  - This module matches the URLs returned by 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.

(ii) Candidate Discovery Module
  - This module finds the SubParent URLs after issuing Parent URL on to browser interface.
  - It downloads the SubParent URL to get the child pages.
  - It extract (anchor text, URL) pair to obtain child URLs.
  - It matches the Parent URLs with Child URLs and obtains candidate synonym 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.

(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 for 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.

(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.

Figure 1. Basic Architecture of Proposed Methodology

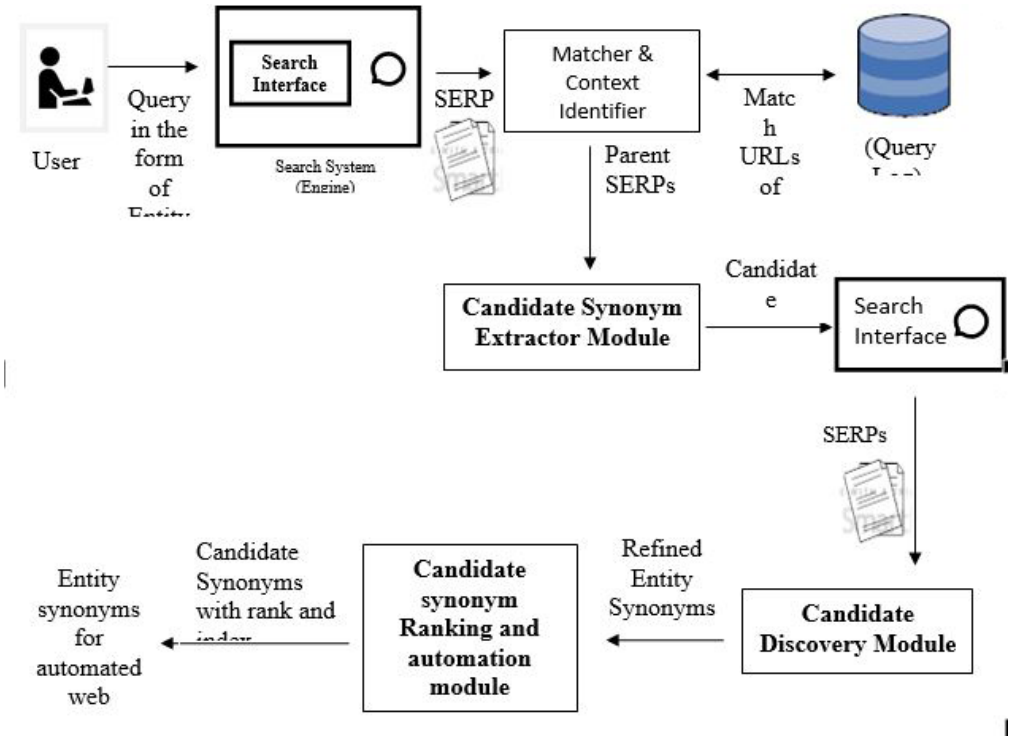
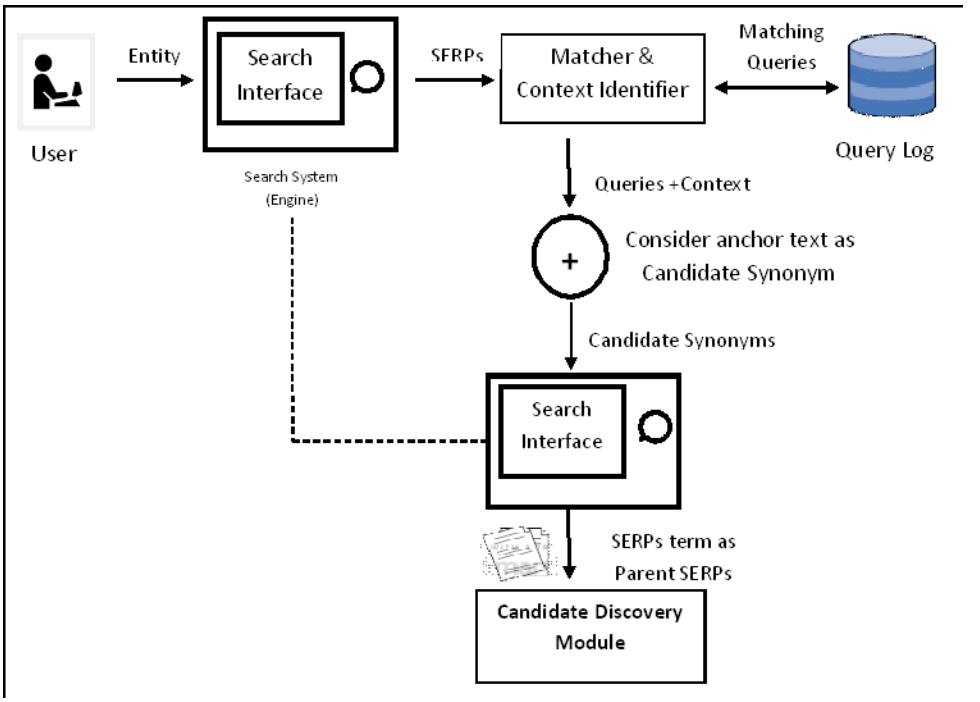


Figure 2. Architecture of Candidate Synonym Extractor Module



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

5.1 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 when the URL returned by search engine is same as the URL already present in the query log. //Let it be $\{E_1,E_2,E_3,\ldots\ldots,E_n\}$
   2.2 Extract the URL of page p in order to retrieve context from them.
   2.3 Find context using snippets and title related to p. //Let it be $\{C_1, C_2, C_3 \ldots C_n\}$
   2.4 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 $\{E_1C_1, E_2C_2 \ldots\ldots\ldots E_2C_1, E_2C_2 \ldots\ldots\ldots E_nC_n\}$ considered as entity synonyms.
3. Submit each candidate synonym as a query to the search interface

Figure 3. Entity Candidate Discovery Module

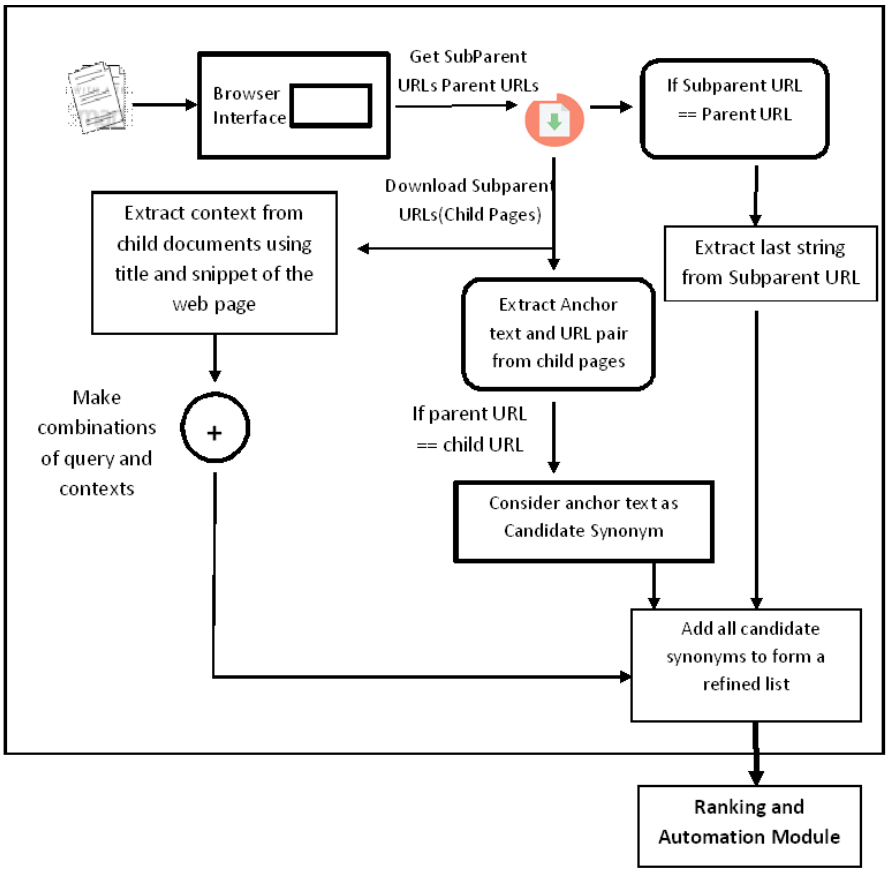
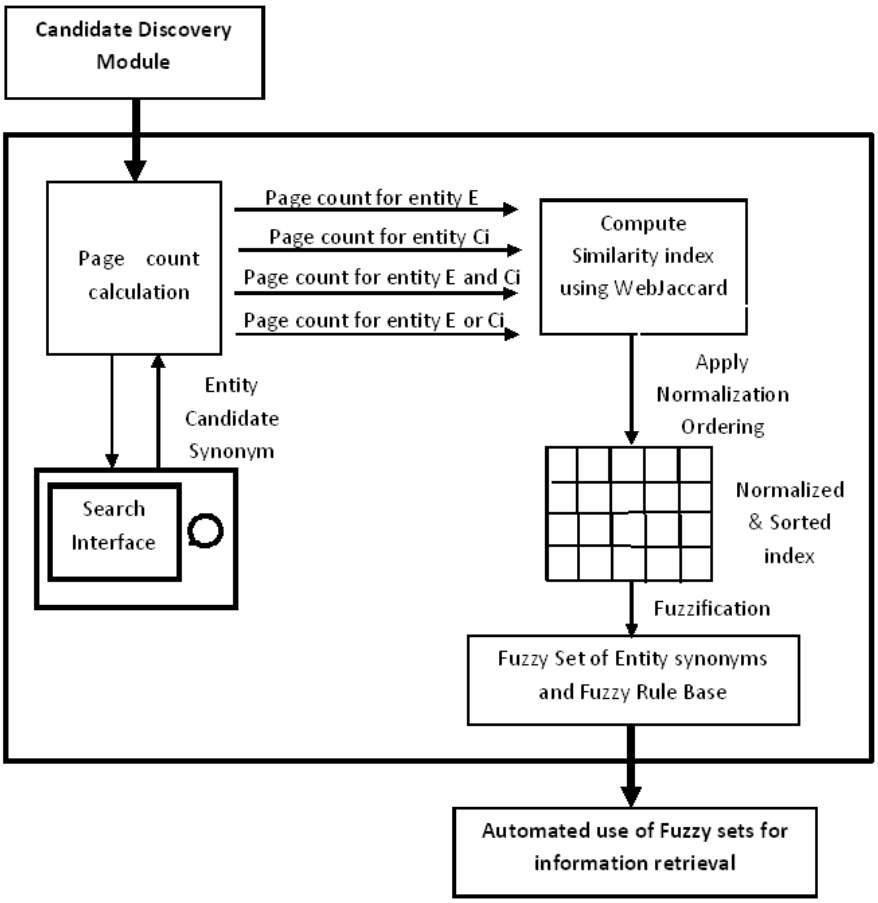


Figure 4. Candidate Synonym Ranking and Automation module



4. newSERPs = SearchEngine(EiCi)
5. Refined_Entity_Synonyms=CandidateDiscovery(newSERPs)
6. Return(Refined_Entity_Synonyms)
7. End.

```
Algorihm: CandidateDiscovery(newSERPs)
// Entity Candidate Discoverer Module
Input: newSERPs (search result pages returned by entity_synonym_
extractor)
Output:Refined_Candidate_synonym
```

1. Treat URLs corresponding to newSERPs as Paresnt URLs(PU's)
2. For Each URL U_i ε PU_i Submit U on browser interface and pick 10 subURLs corresponding to U //call the subURLs 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

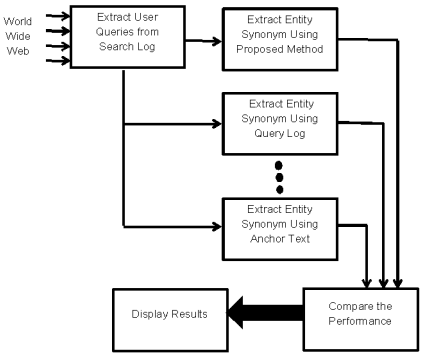
2.1.2 Extract anchor text and URL pair from each child document and store it in a map with attributes as child URL and its corresponding anchor text.
2.2 If PU_i== child URL
2.2.1 Extract anchor text and combine them to the list of candidate synonyms (i.e. CS_i).
2.3 If (SPUs)== (PU_i)
2.3.1 Extract last string from Subparent URL and add them 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 U CS_j U CS_k
5. //Ranking Candidate Synonyms
   Retrieve the page counts for *E (NE)*
   For each C*i* Î Refined_CS do
   Retrieve the page counts for *E* and C*i (NEci)*
   Retrieve the page counts for C*i (Nci)*
   Compute WebJaccard *(E,Ci) = NEci /(NE + NCi − NEci)*
   End for

6. THE EXPERIMENT

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. Figure 5 Shows the overview of the experimental setup.

Figure 5. Overview of the Experimental Setup.



7. RESULTS & DISCUSSION

Table 1 shows the candidate synonyms generated through different approaches including the proposed one. It can be seen that the result set in the proposed approach is much richer than the others. Also,

similarity index (SI) has been shown with entity synonym computed on the basis of Web Jaccard coefficient as described in the algorithm.

The proposed approach can obtain more entity synonyms as compared to conventional approaches which improve the user experience to large extent as seen in Table 1. To enumerate the effectiveness of the proposed system, precision is used as a standard metric.

The precision is calculated for each approach based on the number of relevant results 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 2 and Figure 6.

While comparing with the other existing approaches as shown in Table 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.

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 are 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 approach.

8. HALLMARK 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 can 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, do not cater to the need of dynamic and unstructured nature of WWW (World Wide Web).

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 titles of the webpage. The algorithm also introduces a new approach of finding candidate synonyms from trailing part of sub parent URL.

When one talks of entities, it reminds him/her of Google's knowledge graph. Google's knowledge graph is vast containing so many entities and their relationships. We have made a small entity knowledge graph for representing entities and their candidate synonyms similar to the mechanism used by Google to enhance its search engine's search 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 are shown below.

Figure 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.

Figure 8 shows relationship between two entities and their candidate synonyms.

Table 1. Candidate Synonyms Generated through Query Log, Inbound Anchor Text, Anchor Text+ Context and Proposed Methodology.

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. 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

continued on following page

Table 1. Continued

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
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

9. APPLICATION OF RESULTS

The computed index results are normalized to the 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 autosuggestion, query expansion, query auto replacement, etc. thereby enriching the users' search experience in the order as given in Table 3.

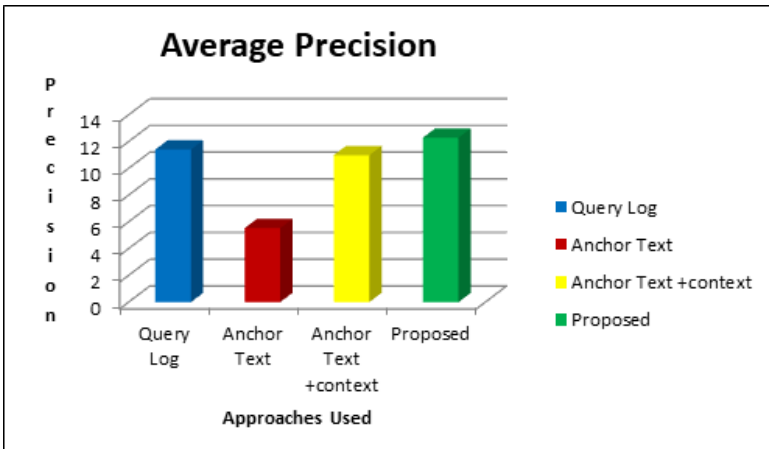
10. CONCLUSION AND FUTURE SCOPE

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

Table 2. Number of relevant results 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

Figure 6. Average Precision for Conventional and Proposed Approach



be used for automated search processes by the search engine 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.

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

Figure 7. Knowledge graph for the Entity indiatimes

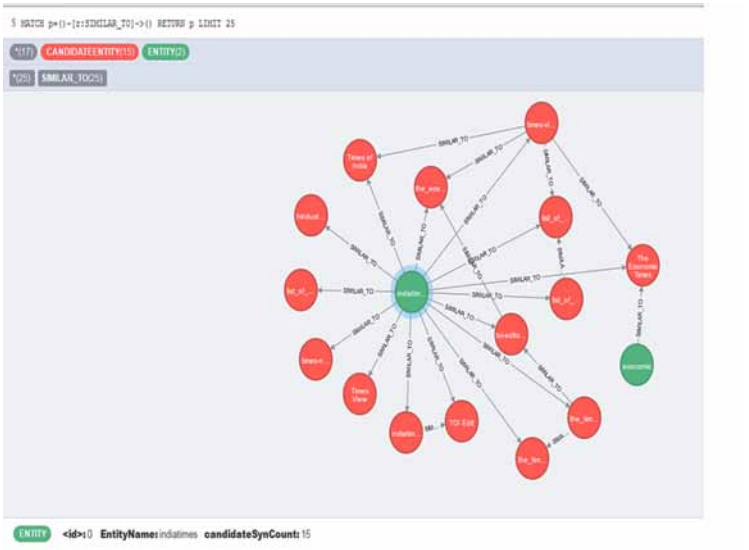


Figure 8. Knowledge graph showing the relationship between two entity synonyms and their candidate synonyms

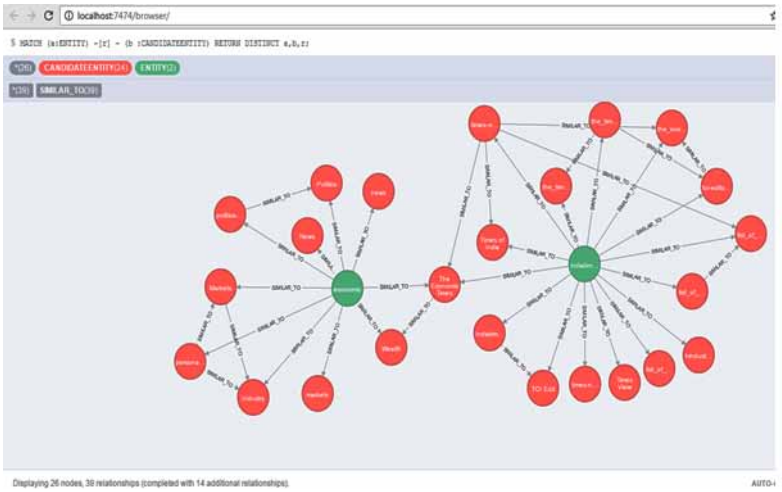


Table 3. (Creation of Fuzzy Sets)

Name of the fuzzy set	Support	Rank	Usage
Excellent entity synonym	[0.80,1.00]	1	Auto Query suggestion, query expansion, query replacement
V. 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

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

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 extracting information from unstructured data. To extend the work, more parameters can be considered to improve the quality of synonym discovery accuracy.

FUNDING AGENCY

Publisher has waived the Open Access publishing fee.

REFERENCES

Bhattacharya, I., & Getoor, L. (2007). Collective Entity Resolution in Relational Data. *ACM Transactions on Knowledge Discovery from Data*, *1*(1), 1–36. doi:10.1145/1217299.1217304

Benjelloun, O., Garcia-Molina, H., Menestrina, D., Su, Q., Whang, S. E., & Widom, J. (2009). Swoosh: A Generic Approach to Entity Resolution. *The VLDB Journal*, *18*(1), 255–276. doi:10.1007/s00778-008-0098-x

Bollacker, K. D., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. SIGMOD.

Chakrabarti, K., Chaudhuri, S., Cheng, T., & Xin, D. (2012). A framework for robust discovery of entity synonyms. *KDD: Proceedings / International Conference on Knowledge Discovery & Data Mining. International Conference on Knowledge Discovery & Data Mining*, *12*, 1384–1392. doi:10.1145/2339530.2339743

Chaudhuri, S., Ganti, V., & Xin, D. (2009). Exploiting Web Search to Generate Synonyms for Entities. *Proc. 18th Int'l Conf. World Wide Web (WWW)*. doi:10.1145/1526709.1526731

Chaudhuri, S., Ganti, V., & Xin, D. (2009). Mining document collections to facilitate accurate approximate entity matching. *Proc. VLDB Endow., 2*(1), 395–406. doi:10.14778/1687627.1687673

Cheng, T., Lauw, H. W., & Paparizos, S. (2012). Entity synonyms for structured web search. *IEEE Transactions on Knowledge and Data Engineering*, *24*(10), 1862–1875. doi:10.1109/TKDE.2011.168

Christian, B., & Kjetil, N. (2010). Extracting Named Entities and Synonyms from Wikipedia. *In advanced information networking and applications*, 1300-1307.

Dey, D., Sarkar, S., & De, P. (2002). A Distance-Based Approach to Entity Reconciliation in Heterogeneous Databases. *IEEE Transactions on Knowledge and Data Engineering*, *14*(3), 567–582. doi:10.1109/TKDE.2002.1000343

Dong, X., Halevy, A., & Madhavan, J. (2005). Reference Reconciliation in Complex Information Spaces. *Proc. ACM SIGMOD Int'l Conf. Management of Data*.

Hu, J., Fang, L., Cao, Y., Zeng, H., Yang, Q., & Chen, Z. (2008). Enhancing Text Clustering by Leveraging Wikipedia Semantics. *Proc. 31st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval*. doi:10.1145/1390334.1390367

Malekian, A., Chang, C., Kumar, R., & Wang, W. (2008). Optimizing Query Rewrites for Keyword-Based Advertising. *Proc. Ninth ACM Conf. Electronic Commerce (EC)*.

Kathuria, M., Nagpal, K. C., & Duhan, N. (2018). Creation of Entity Synonyms Dictionary and its Usage for Query Reformulation: A Review. *Journal of Emerging Technologies and Innovative Research, 5*(8), 1185-1190.

Pei, S., Yu, L., & Zhang, X. (2020). Set-aware Entity Synonym Discovery with Flexible Receptive Field. *IEEE Transactions on Knowledge and Data Engineering*. Advance online publication. doi:10.1109/TKDE.2021.3087532

Shen, J., Lyu, R., Ren, X., Vanni, M., Sadler, B., & Han, J. (2019, July). Mining entity synonyms with efficient neural set generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 249–256. doi:10.1609/aaai.v33i01.3301249

Shen, J., Qiu, W., Shang, J., Vanni, M., Ren, X., & Han, J. (2020). *Synsetexpan: An iterative framework for joint entity set expansion and synonym discovery.* arXiv preprint arXiv:2009.13827. 10.18653/v1/2020.emnlp-main.666

Srikantiah, K. C., Roopa, M. S., Krishna, K. N., Tejaswi, V., Venugopal, K. R., & Patnaik, L. M. (2013). Automatic Discovery of Synonyms from the Web based on Inbound Anchor Text. ICDMW, 130–139.

Stark, M. M., & Riesenfeld, R. F. (1998). Wordnet: An electronic lexical database. In *Proceedings of 11th Eurographics Workshop on Rendering*. MIT Press.

Strube, M., & Ponzetto, S. P. (2006). Wikirelate! Computing Semantic Relatedness Using Wikipedia. *Proc. 21st Nat'l Conf, Artificial Intelligence*.

Thada, V. & Jaglan, V. (2016). 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*(4).

Xiang, R., Cheng, T., & Xiang, R. (2015). Synonym Discovery for Structured Entities on Heterogeneous Graphs. In *International World Wide Web Conference Committee (IW3C2), WWW 2015 Companion*. ACM.

Yang, Y., Yin, X., Yang, H., Fei, X., Peng, H., Zhou, K., Lai, K., & Shen, J. (2021). *KGSynNet: A Novel Entity Synonyms Discovery Framework with Knowledge Graph.* arXiv preprint arXiv:2103.08893. 10.1007/978-3-030-73194-6_13

Zhang, C., Li, Y., Du, N., Fan, W., & Yu, P. S. (2018). *Entity Synonym Discovery via Multipiece Bilateral Context Matching.* arXiv preprint arXiv:1901.00056.

Mamta Kathuria is currently working as an Assistant Professor in J.C. Bose University of Science & Technology, YMCA, Faridabad and has fourteen years of teaching experience. She received her Master in computer application from Kurukshetra University, Kurukshetra in the year 2005 and M.Tech from MDU, Rohtak in 2008. She completed her Ph.D. in Computer Engineering in 2019 from J.C. Bose University of Science and Technology, YMCA. Her areas of interest include Artificial Intelligence, Web Mining, Computer Vision, and Fuzzy Logic. She has also published over 50 research papers in reputed International Journals including SCI, Scopus, ESCI, UGC and Conferences.

Chander Kumar Nagpal received Ph.D (Computer Science) from Jamia Milla Islamia, New Delhi. He is currently working as a professor in YMCA University of Science & Technology and has thirty three years of teaching experience. He has published two books. He has published many research papers in reputed international journals such as IEEE transaction on software reliability, Wiley STVR, Springer Multimedia tools and applications, Journal of CSI, etc. His academic interests include Web Mining, Information Technology, Predictive Analytics, and Soft Computing.

Neelam Duhan received her B.Tech. in Computer Science and Engineering with Honors from Kurukshetra University, Kurukshetra, and M.Tech with Honors in Computer Engineering from Maharshi Dayanand University, Rohtak in 2002 and 2005, respectively. She completed her Ph.D. in Computer Engineering in 2011 from Maharshi Dayanand University, Rohtak. She is currently working as an Associate Professor in the Computer Engineering Department at YMCA University of Science and Technology, Faridabad, and has experience of about 15 years. She has published over 50 research papers in reputed International Journals and International Conferences. She is the recipient of AICTE Visvesvaraya best teacher award- 2021 at the national level. Her areas of interest are databases, search engines and web mining.