International Journal of Future Generation Communication and Networking
Vol. 10, No. 11 (2017), pp.19-36
http://dx.doi.org/10.14257/ijfgcn.2017.10.11.03

Discovery of Entity Synonym Using Anchor Text and URLs

Mamta Kathuria[1], Anurahda Singh[2], C. K. Nagpal[3] and Neelam Duhan[4]

[1]Assistant Professor, YMCA University of Science & Technology, Faridabad (India)
[2]Student (M.Tech), YMCA University of Science & Technology, Faridabad (India)
[3]Professor, YMCA University of Science & Technology, Faridabad (India)
[4]Assistant Professor, YMCA University of Science & Technology, Faridabad (India)
[1]mamtakathuria@ymcaust.ac.in, [2]anuradhasngh13@gmail.com, [3]nagpalckumar@rediffmail.com, [4]neelam.duhan@gmail.com

Abstract

In the current scenario, the web queries have become more and more pin-pointed so as to find results relating to specific entity in a specific context of time, place, etc. For example, information pertaining to a movie-show, a particular train, newspaper of a particular date, performance of a particular stock etc. All these references associated with a particular entity are known as entity references. The problem with these references is that they vary with the heterogeneous contexts of the web and one may not be getting the required answers to his/her query owing to these varied entity references known as entity synonyms. These entity synonyms cannot be handled through lexical resources like WordNet [1]. Therefore, every search engine will have to create its own mechanism for finding the entity synonyms of a particular entity in order to properly answer the users' queries, the process being known as entity resolution. In recent past, many researchers have tried to devise the mechanisms to generate the entity synonyms. This paper is also an effort in this direction and creates a rich set of entity synonyms for a given entity using inbound anchor text and URLs.

***Keywords**: Entity, Candidate Entity Synonym, Web Query, Inbound Anchor text, Entity Synonym Extraction*

1. Introduction

With the growth of the web, users have been making diverse forms of queries relating to a variety of domains concerned with daily life issues. These queries associated with products, brands, recipes, weather forecast, show timings, quotes for various products *etc*. are being searched by common users to accomplish their daily needs. These searches related to entities can be best sorted out from the latest product catalogs and associated databases if the references are specific. However, if the references are general and refer to common entities, then product catalogs and databases may not be available. When the entities are well known, the synonyms can be found by the usage of sources like Wikipedia [2] and FreeBase [3]. For instance, the Bhabha Atomic research Center may be referred as Bhabha Institute, BARC, Atomic Energy Center, and Nuclear Energy Center *etc*. However, for the common entities these online resources may not work.

The problem with these generic entities is their multiple types of references to the same entity due to different creators of the web pages. For example, a paper like The Hindustan Times may be referred as The HT, HT, The Hindustan, The Hindustan Times Today, and

Received (May 15, 2017), Review Result (August 31, 2017), Accepted (September 20, 2017)

Hindustan Times *etc*. The movie Bahubali-The Conclusion is also referred to as Bahubali-2. Thus, there is a requirement to devise a mechanism to find entity synonyms for the common entities to enable the search engines to provide a rich search experience to its users.

The best possibility to find various varieties of entity synonyms available on the web is through the exploration of the web and to find the set of possible candidate references which can be further pruned. The paper considers such search efforts carried out in the recent past and proposes a novel technique that creates a rich set of entity synonyms. The proposed work will help in improving search relevance, enrich users' experience, query auto-suggestion, creation of entity dictionary, recommendation system and e-discovery *etc*.

The rest of the paper is organized as follows: Section 2 contains the formal description of the concepts used in this paper. Section 3 talks about web based empirical methods available in the literature to find out the entity synonym set for a given input string. Section 4 defines problem and objectives set for the work. Section 5 contains details of proposed methodology for entity resolution that provides better and enhanced results compared to its predecessors. Section 6 discusses the usage of results. Section 7 talks about the conclusion and future scope of the work.

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, we take the help of the following concepts:

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 ε S and the second one being the entity domain E_X. Then the function $F(s, E_X)$-> e E maps the string s to a single or a set of entities in the global domain E which is a superset of Ex , 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:

$$Ws = \{w \in S \mid F(s, E_X) = F(w, E_X)\}$$

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.

The subsequent section talks about such empirical methods as used by the various researchers along with their merits and demerits.

3. Literature Survey

With the time, growth of the web round the globe has made it a universally accepted information resource. This vast ocean of knowledge contains wide variety of references to the each entity, made by heterogeneous types of content creators and the web searchers. These heterogeneous references, if not properly taken care of, shall lead to inadequate supply of information to the web searcher despite the availability of the information. A need, therefore, has been long felt to cover this gap by grouping the references leading to same common entity into a set known as the set of entity synonyms and the process is referred to as entity resolution. The primitive approach in this regard has been the use of abbreviations, acronyms, Wikipedia and Freebase references. All these mechanisms have not been able to cover the web as a whole and its heterogeneity. Thus only solution for creating these set of references (entity synonyms) must include the use of web-ocean as source and fishing out the references. The literature of past ten years contains many such efforts. This section covers to these efforts and discusses their relative merits and demerits.

Pinky Paul *et al.,* [4] have presented a study about entity search engine. They have discussed the basic architecture of entity search engine as shown in Figure 1. The approach is static and involves the possible basic modules of an Entity Search Engine. These modules accomplish the task of entity candidate extraction from the web, categorization, assignment of relevance and ranking. The architecture is quite general and provides an overview of the process.

The different approaches used in the entity resolution may differ in the method of entity extraction process which may be offline or online. The result computation can also be based upon some intermediate approach like pseudo document or reference table *etc*. The synonym generation process can also be static or dynamic. We take up different approaches in the above context.

Surajit Chaudhuri *et al.,* [5] have proposed a method to find correlation between a candidate string and an entity in order to overcome the shortcomings of string-based similarity measures that does not reflect the common knowledge that users generally provide for the candidate string in the question. They proposed new document-based similarity measures to calculate the similarity in the context of many documents containing the candidate string. The main disadvantage of their approach is that they have performed computations on a small set of relevant documents containing the given candidate string to measure the correlation.

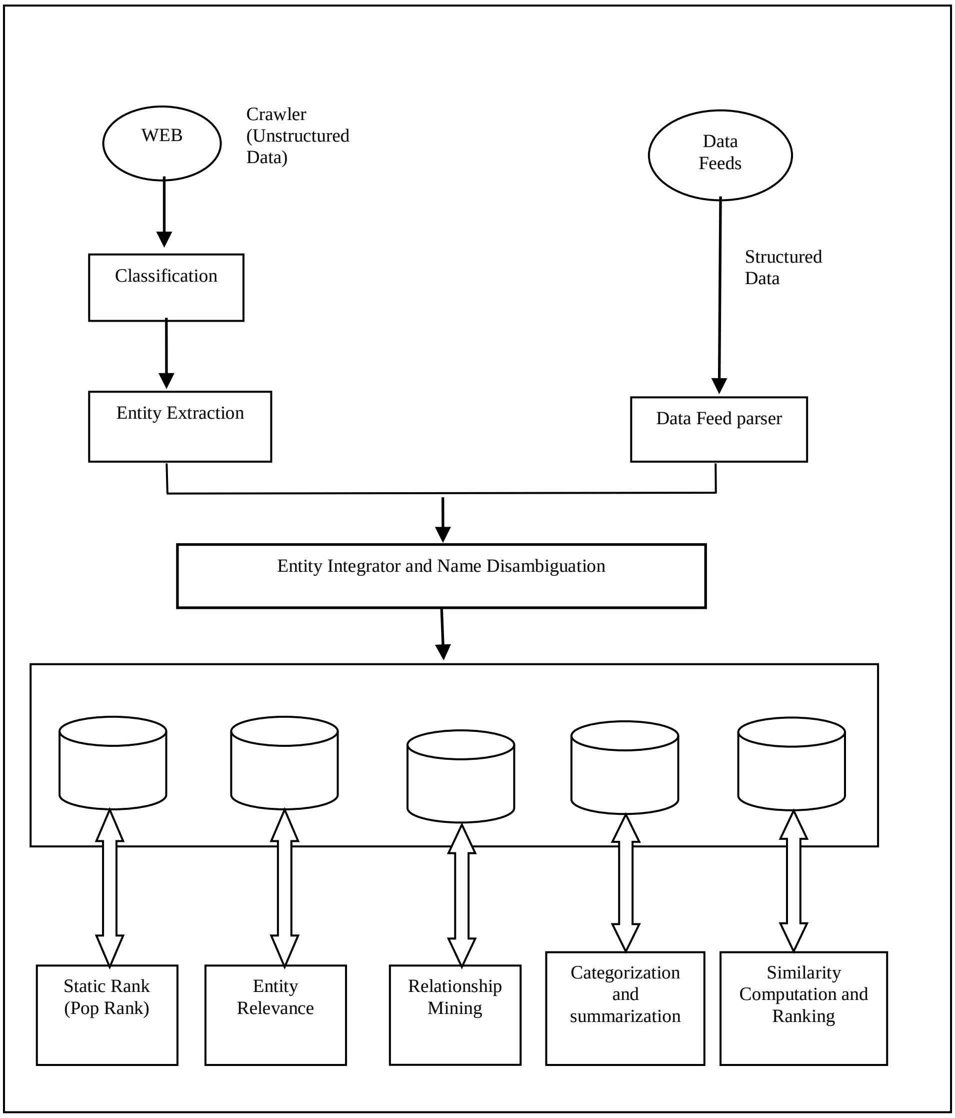
WEB
Crawler (Unstructured Data)
Data Feeds
Classification
Structured Data
Entity Extraction
Data Feed parser
Entity Integrator and Name Disambiguation
Static Rank (Pop Rank)
Entity Relevance
Relationship Mining
Categorization and summarization
Similarity Computation and Ranking

Figure 1. General Architecture of Entity Search Engine

Tao Cheng *et al.* [6] have proposed an offline fully automated data-driven algorithm called as Identifying Normalization Entity Synonym. This algorithm mines queries where a variety of keywords have been used to refer to the same web pages and generates an expanded set of equivalent string called entity synonyms for each keyword. This architecture comprises of three modules: Candidate generation, Candidate Selection and noise cleaning. This algorithm works well for structured data and covers the structured web queries with good precision. However, this technique is not applicable to dynamic and unstructured data.

Hamid Mousavi *et al.* [7] proposed a new technique to generate context aware synonyms for the entities and attributes. They proposed the Context-aware synonym Suggestion System (CS3) which learns synonyms from text by using NLP-based text mining technique called SemScape. The architecture comprises of four parts:

- Integrating existing knowledge bases by converting them into a common internal representation and storing the integrated one in another knowledge store called as Integrated Knowledge Base (IKB) store.

- Further addition of facts from free text to IKB store.
- Large-corpus generation of context-aware synonyms that can be used to remove inconsistencies in IKB store.
- Incompleteness resolution in IKB store by using the synonyms generated in above step.

The main disadvantage of the proposal is its static approach. The knowledge bases are bound to change over time and need to be refreshed for fresh synonyms. Moreover, the approach does not tackle dynamic nature of web.

L. Jiang *et al.* [8] have proposed GRIAS, a Graph-based framework for discovering entity aliases motivated by the entity relationships collected from both the structured and unstructured data. The goal of this paper is to accurately identify entity aliases, especially the long tail ones from the unstructured data. The graph-based similarity is calculated between an entity and its alias candidates chosen by candidate selection method using entity relationship graph. The approach is quite static.

Tao Cheng *et al.* [9] have proposed an offline data-driven bottom up approach that mines query logs. They have made a fully automated solution that works on structured data to find reference entity synonyms for various entities. To achieve this, they have used a multi-step data driven approach that relies on query and click logs. The solution works by retrieving relevant web page URLs that can act as good representatives of the entity. Then, by following the URL-query click graph, they identified a list of URLs from the URLs identified earlier. The query words corresponding to this URL list will act as true synonyms by inspecting click patterns and click volume on a large subset of such urls. But, since this approach is offline and structured, it cannot handle unstructured and dynamic data evolving over web.

Surajit Chaudhuri *et al.* [10] have created large reference entity tables, based upon the concept initially proposed by [11], by mining entity variant from different documents. An input string is matched with the references available in reference table. If a match is found, the content of the table serve as entity synonyms.

Yanen Li *et al.* [12] have proposed a comprehensive approach towards entity resolution based upon syntactic patterns, query entity clicks and distribution similarity. They have focused in the fact that an entity synonym must comply with the associated context. The approach is offline and based upon the clustering model. The merit of the system is its ability to handle the heterogeneity but the scheme suffers from the drawback of static approach.

Kaushik Chakrabarti *et al.* [13] have proposed a general framework for discovering entity synonym based upon two novel similarity measures called as Psuedo Document Similarity and Query Context Similarity to create a set of entity synonyms. The approach is offline and dependent upon query logs.

Harada *et al.* [14] have proposed an approach called NEXAS (Named Entity extraction and Association Search) that finds authoritative people from web by associating a web page through identification of its real world entities. It determines nearly all relevant entities by considering top ranked web pages from SERPs. The approach is limited to people only and can easily be better substituted through references like Wikipedia. Similar arguments can be given for the work done by Kalashnikov *et al.* [15] and Bollegala *et al.* [16].

Xiang Ren *et al.* [17] have proposed a general heterogeneous graph-based data model that encodes problem of structured view of each entity by capturing three key concepts Synonym candidate, Web page, Keyword and different types of interactions between them. They proposed a graph based approach for ranking entity synonyms and demonstrated a closed-form optimal solution for outputting entity synonyms scores. They adopted structured view of each entity by taking into consideration other important structured attributes along with string name. They make use of sub-queries, tailed synonyms and tailed web pages for harvesting more synonyms. The approach is offline and suffers from a priori the requirement of candidate synonyms.

Roi Blanco *et al.* [18] have proposed a recommendation engine named as Spark that links initial query of user to an entity within knowledge base. It provides ranking of the related entities. The proposed system extracts several signals from a variety of data sources that includes user sessions, twitter and flicker using various cluster of computers running Hadoop. The extracted signals are combined with a machine learned ranking model that produces final recommendation of entities to user queries. Spark has been currently used in Yahoo search result pages.

Srikantiah *et al.* [19] have 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 paper.

Most of the above approaches are offline and static, thus cannot cater to the need of dynamically expanding web. The approach presented in this paper, uses the dynamic context created through anchor text, context and trailing part of the dynamically generated URL. The subsequent section defines the problem and the objectives to be fulfilled by the proposed work.

4. Problem Definition and Objectives

The proposed work aims to generate a rich-set of entity synonyms for a given entity string, says, through online dynamic approach using the anchor text, context and trailing part of URL; and to rank the synonyms based on co-page count basis.

The objectives to be achieved by the proposed approach are:

- To create an automated mechanism for generating synonyms that can be used for auto-suggestion, auto-replacement, query expansion *etc*.
- The method should be usable for commonly used entities in addition to well-known entities.
- The process should be dynamic and online.
- The entity synonyms set should be rich as compared to the previous approaches.

5. Proposed Approach for Entity Synonyms Discovery

The proposed work involves the usage of Search Engine Result Pages (SERPs) for extracting candidate entity synonyms, combines contexts with the input entity string thereafter uses anchor text of the downloaded web pages and the trailing part of the sub parent URLs to create a set of entity synonyms. The whole work of extracting synonyms is dividing into two parts

- Extraction of candidate entity synonyms
- Ranking of these synonyms based on page-count measure

The algorithm first uses the input entity to generate SERPs and URLs corresponding to those pages are collected in a list of parent URLs (PUs). Next, these parent URLs 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 are 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 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 three things:

- From child map in case of match
- From trailing part of sub parent url
- From combination of query and context obtained from child documents

The detailed process is shown through a flowchart as shown in Figure 2.

The page count of an entity E is an estimate of the number of pages that contain the entity sting E. The candidate entity synonyms are ranked using similarity measure which in turn uses page counts of E only, candidate synonym ci only and both E, ci. A similarity measure or similarity function is a real-valued function that quantifies the similarity between two objects. It represents the similarity between two objects or words.

Let NE be the page count corresponding to input entity E only, Nci be the page counts for a candidate synonym ci only and NEci be the page counts for both entity string E and a candidate synonym ci. The proposed works uses WebJaccard similarity measure to accurately measure the relevance between E and ci, which is further used to rank the candidate synonyms using page counts.

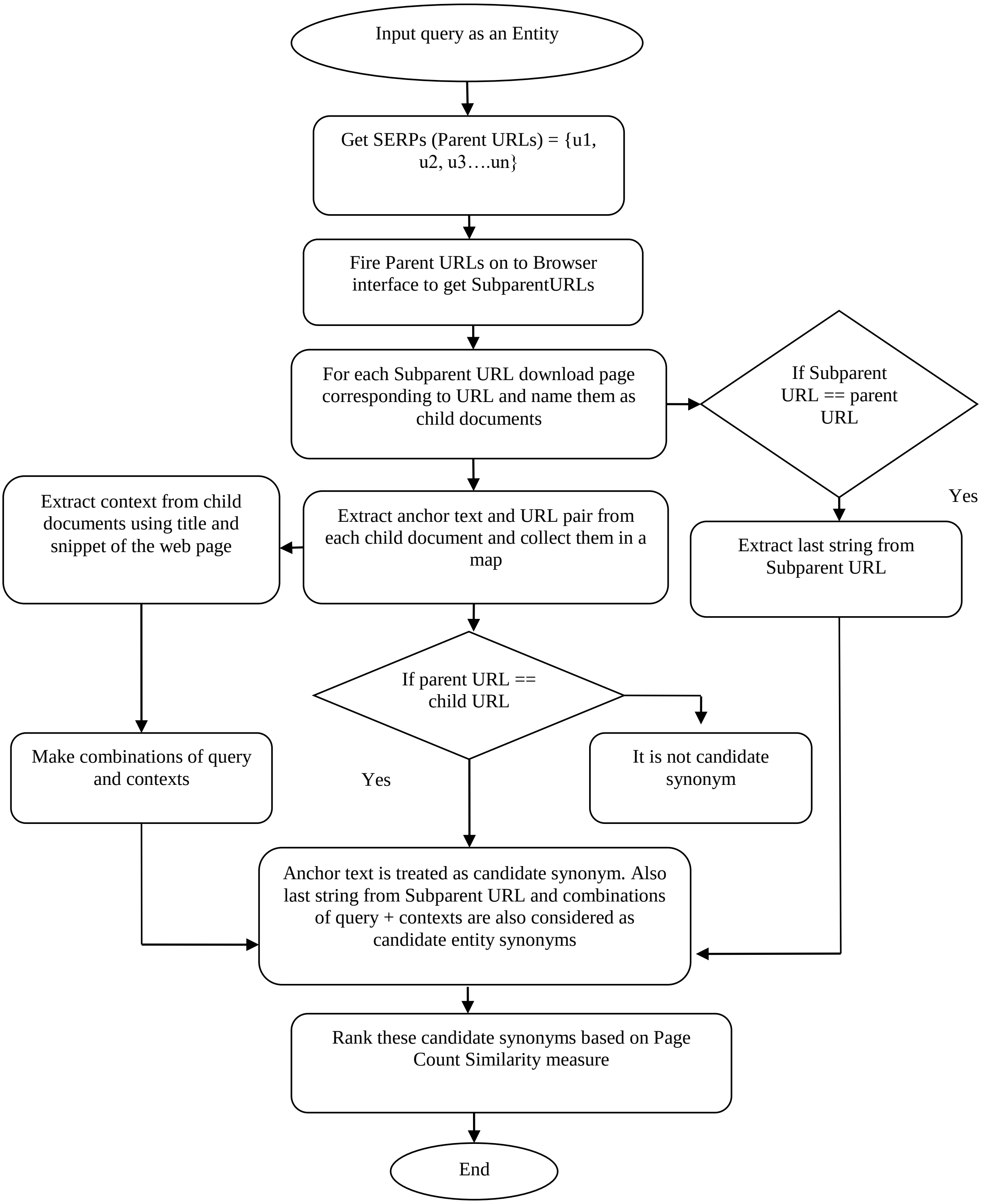
Input query as an Entity
Get SERPs (Parent URLs) = {u1, u2, u3….un}
Fire Parent URLs on to Browser interface to get SubparentURLs
For each Subparent URL download page corresponding to URL and name them as child documents
If Subparent URL == parent URL
Yes
Extract last string from Subparent URL
Extract context from child documents using title and snippet of the web page
Extract anchor text and URL pair from each child document and collect them in a map
If parent URL == child URL
Yes
It is not candidate synonym
Make combinations of query and contexts
Anchor text is treated as candidate synonym. Also last string from Subparent URL and combinations of query + contexts are also considered as candidate entity synonyms
Rank these candidate synonyms based on Page Count Similarity measure
End

Figure 2. Flowchart of Proposed Algorithm

The details of the process have been described through the following algorithm.

Algorithm (E, CS)
Input**:** Entity String E.
Output**:** List of Candidate Synonyms for E ranked using page count measures
//Generation of Entity Candidate Synonyms

]

```
1. Submit the input query E on to the search interface to get Search Engine Result
   Pages (SREPs) from search engine.
1.1. Extract all Urls from SERPs obtained above into a list of parent Urls (i.e PU)

2. For each URL $u_j \epsilon PU$ do
2.1. Get all pages that contain URL $u_j$ into another list called as sub parent urls.
Thus, for each $u_j$ there will be a list called as $SPU_j$.
2.2. Sub parent list (SPU) = $\cup_j SPU_j$

3. For each URL $su_k \epsilon SPU$ do
3.1. $childDocument_k$ = Downlaod the web page corresponding to URL $su_k$
3.2. childDocuments = $\cup_k childDocuments_k$

4. For each $childDocument_i \epsilon childDocuments$ do
4.1. Collect all anchor text and URL pair ($u_j$, $a_j$) from $childDocument_j$ into a hash map
   (i.e $CM_j$)
4.2. Make union of all the pair collected above into a single map of anchor text and its
   corresponding URL. (i.e CM(child map) = $\cup_j CM_j$)
4.3. Also extract context from each child document and collect them into a list of
   contexts. Context of the web page is extracted using title and snippet of the web
   page.

5. For each $ci \epsilon Context$ do
5.1. Combine the initial entity string E with context $ci$
5.2. Maintain the above combination of entity string and context into a list called as
   QC(Query context)

6. For each $u_j \in PU$ do
For each URL $su_k \in SPU$ do
If ($u_j == su_k$) then
6.1. trailingPartofURL = Extract last string from subparent url
6.2. Collect all strings obtained into a list of synonym obtained from trailing part of
   url called as TPU

7. For each url $u_i \epsilon PU$ do
For each <url, anchortext> pair $u_j a_j \epsilon CM$ do
If ($u_i == u_j$) then
7.1. Collect the anchor text associated with url $u_j$ into a list of Candidate Synonyms
   (CS)

8. Union synonyms obtained from child documents, <query + context> and trailing
   part of sub parent url into one list of candidate synonyms
8.1. Thus, CS = ∪ TPU ∪ QC

9. //Ranking of Entity candidate synonyms
Retrieve the page counts for $E (NE)$
For each candidate synonym $ci \in CS$ do
Retrieve the page counts for $E$ and $ci (NEci)$
Retrieve the page counts for $ci (Nci)$
Compute WebJaccard $(E, ci) = NEci /(NE + Nci - NEci)$
End for
```

Table 1 shows the results of the proposed approach and its comparison with [19].

Table 1. Comparison Table

Sr. No .	Entity	[19] Results	SI	Results of Proposed Algorithm	SI	List of Context used in Proposed Methodology
1	inception	Inception - Wikipedia	0.013	inception will	1.000	tv, arrival, wiki, overuse, while, title, -LRB-, imdb, inception/nostalgia, will, cillian, film, premiere, spot, Christopher
				inception title	0.525	
				inception arrival	0.395	
				inception spot	0.294	
				inception tv	0.275	
				inception film	0.237	
				inception Christopher	0.142	
				Inception (2010) – IMDb	0.049	
				inception overuse	0.018	
				inception wiki	0.013	
				Inception – Wikipedia	0.013	
2	Nobel prize	"The Nobel Prize in Physics"	0.048	nobel-prize	0.984	year, ernst, chandrasekhara, foundation, saal, prize, norwegian, riksbank, assembly, venkata, winners, karl, maathaus, parody, prize, curie, prizes, peace, sverige, step, Einstein
				nobel prize winners	0.759	
				nobel prize foundation	0.166	
				"The Nobel Prize in Physics"	0.048	
				nobel prize peace	0.032	
				nobel prize assembly	0.019	
				nobel prize Einstein	0.018	
				nobel_prize_in_physics	0.014	
				nobel prize curie	0.013	
				nobel_prize	0.007	
				"Nobel Prize | award"	0.000	
3	Bard of Avon	The Bard of Avon, William Shakespeare	0.054	Bard of Avon shakespeare	0.896	shakespeare, definition
				why-is-shakespeare-called-the-the-bard-of-avon	0.826	
				bard-of-avon-the-story-of-william-shakespeare	0.806	
				bard--of—avon	0.089	
				The Bard of Avon, William Shakespeare	0.054	
4	onocrotalus	HBW Alive	0.006	onocrotalus pelecanus	0.895	pelecanus, factsheet, pelican, download, base
				pelecanus_onocrotalus	0.797	
				onocrotalus download	0.643	

				onocrotalus pelican	0.588	
				onocrotalus base	0.434	
				great-white-pelican-pelecanus-onocrotalus	0.365	
				onocrotalus factsheet	0.022	
				HBW Alive	0.006	
5	lala lajpat rai	Lala Lajpat Rai - Cultural India	0.260	lala lajpat rai university	0.857	image, period, university, raus, lala, lajpat, biography, thing, newsome
				lala lajpat rai period	0.341	
				Lala Lajpat Rai - Cultural India	0.260	
				lala_lajpat_rai	0.009	
				lala lajpat rai raus	0.004	
				lala lajpat rai newsome	0.002	
6	times of india	India News	0.252	times of india lifestyle	0.802	news, tech, movie, blogs, blog, toi, lifestyle, rss, top, time
				India News	0.252	
				times of india toi	0.013	
7	iupac	IUPAC Nomenclature of Organic Chemistry	0.789	IUPAC Nomenclature of Organic Chemistry	0.789	nomenclature, rname, book, rule, chemistry
		IUPAC Recommendations on Organic & Biochemical Nomenclature, Symbols, Terminology, *etc.*	0.006	iupac name	0.135	
				iupac book	0.090	
				iupac rule	0.087	
				iupac nomenclature	0.047	
				international_union_of_pure_and_applied_chemistry	0.021	
				IUPAC Recommendations on Organic & Biochemical Nomenclature, Symbols, Terminology, *etc.*	0.006	
				iupac_nomenclature_of_organic_chemistry	0.002	
8	anthurium	Caring For Anthurium Growing In The Garden Or Home	0.070	anthurium www	0.770	undying, plants, www, scott, andraeanum, flowers, anthurium, linguifolium, tail, one, è, clarinervium, flower, herbaceous, species,
				anthurium plant	0.626	

				anthurium flower	0.534	plant, information, care
				anthurium flowers	0.526	
				Caring For Anthurium Growing In The Garden Or Home	0.070	
				anthurium herbaceous	0.032	
				anthurium species	0.015	
				anthurium scott	0.014	
				Anthuriums	0.007	
				anthurium andraeanum	0.005	
9	mahatma gandhi	Mahatma Gandhi Biography biography	0.036	mahatma gandhi india	0.721	check, one, photo, k, india, o, godse, gandhus, vinayak, #mahat, gandhi, user
				mahatma gandhi o	0.461	
				Mahatma Gandhi Biography biography	0.036	
				mahatma gandhi vinayak	0.022	
				mahatma gandhi godse	0.019	
				mahatma_gandhi	0.007	
10	harvard university	Harvard.edu	0.397	harvard university link	0.503	college, yard, extension, university, link, university, institution, museum, school, memorial, president
		Harvard College Admissions & Financial Aid	0.124	harvard university president	0.406	
				Harvard.edu	0.397	
				harvard university extension	0.318	
				harvard university institution	0.312	
				harvard university museum	0.190	
				Harvard College Admissions & Financial Aid	0.124	
				harvard university yard	0.040	
				harvard university memorial	0.033	
				harvard_university	0.001	
11	caspian sea	"Caspian Sea"	0.002	caspian sea news	0.493	news, caspian, dissolution, water, sea, monster, sciencecaspian, field, azerbaijani, map
				caspian sea water	0.389	
				caspian sea field	0.190	
				caspian sea map	0.188	
				caspian sea Azerbaijani	0.183	
				Caspian Sea – Wikipedia	0.162	
				caspian sea dissolution	0.097	
				caspian sea monster	0.067	

				Caspian Sea Map, Caspian Sea Location Facts History, Major …	0.017	
				Caspian Sea: Largest Inland Body of Water	0.012	
				caspian_sea	0.011	
				"Caspian Sea"	0.002	
12	moment of inertia	Lecture notes on rigid-body rotation and moments of inertia	0.008	moment of inertia expression	0.482	expression, video, moment, inertia, sewer, physics, cylinder, pipe
				more-on-moment-of-inertia	0.126	
				moment of inertia video	0.098	
				calculating-moment-of-inertia-of-point-masses	0.050	
				moment of inertia sewer	0.043	
				moment of inertia physics	0.040	
				moment of inertia cylinder	0.040	
				moment of inertia pipe	0.037	
				Lecture notes on rigid-body rotation and moments of inertia	0.008	
13	jobs in computer engineering	Computer Engineer	0.384	jobs in computer engineering diploma	0.962	computer, diploma
				computer-science-engineering-jobs	0.921	
				computer-engineer-jobs	0.474	
				Computer Engineer	0.384	
				computer-engineering	0.365	
				Computer Science Engineering	0.152	
14	lokmanya	Bal Gangadhar Tilak	0.067	lokmanya www	0.450	coimbatore, movie, yug, bal, search, bank, society, tilak, www, lokmanyatilak, pronunciation, ek, gangadhar
				lokmanya society	0.176	
				lokmanya tilak	0.171	
				lokmanya Coimbatore	0.165	
				lokmanya gangadhar	0.124	
				Bal Gangadhar Tilak	0.067	
				bal_gangadhar_tilak	0.033	
15	ddlj	Dilwale Dulhania Le Jayenge	0.377	Dilwale Dulhania Le Jayenge	0.378	initialism

				dilwale_dulhania_le_jayenge	0.178	
				ddlj initialism	0.005	
				tt0112870	0.004	
16	isro	Indian Space Research Organisation	0.015	indian_space_research_organisation	0.302	release, orbital, organisation, hq, space, research, gov, notification, recruitment, www, isro, satellite, centres, overview, centre, application, missions, job
				isro release	0.074	
				isro centres	0.070	
				about-isro	0.064	
				isro research	0.063	
				isro space	0.062	
				isro satellite	0.059	
				isro centre	0.057	
				isro job	0.056	
				isro organization	0.055	
				isro gov	0.052	
				isro missions	0.052	
				isro recruitment	0.051	
				isro notification	0.051	
				isro www	0.037	
				isro overview	0.028	
				Indian Space Research Organisation	0.015	
17	gerbera	Gerbera Daisy Care - Tips On How To Grow Gerbera Daisies	0.047	gerbera är	0.284	spain, daisies, fact, ett, gerberas, est, è, gerbera, ist, är, der, jamesonii, un, aurantiaca, hilton, care
				gerbera fact	0.235	
				gerbera Hilton	0.176	
				gerbera ist	0.166	
				gerbera daisies	0.165	
				Gerbera Daisy Care - Tips On How To Grow Gerbera Daisies	0.047	
				gerbera è	0.027	
				gerbera der	0.019	
				gerbera spain	0.014	
				gerbera est	0.013	
				gerbera jamesonii	0.008	
				gerbera_jamesonii	0.007	

18	red fort of india	history behind red fort history of india	0.062	red fort of india moti	0.185	moti, shahjahan
		Red Fort - Wikipedia, the free encyclopedia	0.021	HISTORY BEHIND RED FORT ~ HISTORY OF INDIA	0.062	
				red fort of india shahjahan	0.029	
				Red Fort - Wikipedia, the free encyclopedia	0.021	
				red_fort	0.018	
19	vivekanand	The Life and Teachings of Swami Vivekananda	0.031	vivekanand life	0.157	biographical, bibliography, thought, swami, teachings, life
				The Life and Teachings of Swami Vivekananda	0.031	
				vivekanand teachings	0.030	
				vivekanand swami	0.026	
				vivekanand biographical	0.025	
				vivekanand thought	0.023	
				swami_vivekananda	0.021	
				vivekanand bibliography	0.019	
				stories-swami-vivekananda-life-inspired	0.010	
20	terms used in biology	Glossary of biology - Wikipedia	0.150	Glossary of biology – Wikipedia	0.150	
		Biology Vocabulary: Understanding Common Terms	0.010	biology-vocabulary	0.055	
				Biology Vocabulary: Understanding Common Terms	0.010	

6. Using the Results

The hallmark of proposed scheme is that it not only computes a rich-set of entity synonyms, but also shows the extent of matching with the original entity string. This known extent of matching can be used in the same manner as used in case of alpha-cut in case of fuzzy logic. The designer of the search engine can make decisions about the cut of values for using a synonym for a particular purpose. For example, the lower cut off value for auto-suggestion, auto-replacement and query expansion can be 0.6, 0.75 and 0.9 respectively. The ranks can also be fuzzified by designing the appropriate fuzzy-sets which can be used for automating the query auto-suggestion, replacement and expansion processes through a system like fuzzy-rule based inference system.

7. Conclusion and Future Scope

The proposed work is able to generate a comparatively richer set of entity synonyms than its predecessors and is able to assess the quality of the synonyms as well. The work can be further refined by the augmentation of synonyms retrieved from query log. The relationship between the entity string and the corresponding synonyms can be exhibited using an entity graph through which various possibilities can be explored including the machine intelligence.

References

[1] M. M. Stark and R. F. Riesenfeld, "Wordnet: An electronic lexical database", In Proceedings of 11th Eurographics Workshop on Rendering, MIT Press, 1998J. Breckling, Edition, The Analysis of Directional Time Series: Applications to Wind Speed and Direction, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, vol. 61, **(1989)**.

[2] G. K. Ralf Schenkel and Fabian Suchanek, "Yawn: A semantically annotated wikipedia xml corpus", In GI-Fachtagung fur Datenbanksysteme in Business, Technologie und Web, volume 103 of LNI, **(2007)**.

[3] K. D. Bollacker, C. Evans, P. Paritosh, T. Sturge and J. Taylor, "Freebase:a collaboratively created graph database for structuring human knowledge", In SIGMOD, **(2008)**.

[4] P. Paul and T. George, "Entity Search Engines", In International Journal of Computer Science and Mobile Computing, vol.3, iss. 2, **(2014)**, pg. 877-880

[5] S. Chaudhuri, V. Ganti and D. Xin, "Exploiting Web Search to Generate Synonyms for Entities", In WWW '09 Proceedings of the 18th international conference on World wide web, Madrid, Spain, **(2009)**.

[6] T. Cheng, H. W. Lauw and S. Paparizos, "Entity Synonyms for Structured Web Search", IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 10, **(2012)**, pp. 1862–1873.

[7] H. Mousavi, S. Gao and C. 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, **(2013)**.

[8] L. Jiang, P. Luo, J. Wang, Y. Xiong, B. Lin, M. Wang and N. An. Grias, "An entity-relation graph based framework for discovering entity aliases", In ICDM, **(2013)**.

[9] T. Cheng, H. W. Lauw and S. Paparizos, "Fuzzy Matching of Web Queries to Structured Data", In Data Engineering (ICDE), 2010 IEEE 26th International Conference, **(2010)**.

[10] S. Chaudhuri, V. Ganti and D. Xin, "Mining Document Collections to Facilitate Accurate Approximate Entity Matching", In PVLDB, **(2009)**.

[11] A. Aho and M. Corasick, "Efficient string matching: an aid to bibliographic search", In Comm. ACM, vol. 18, **(1975)**.

[12] Y. Li, B.-J. (Paul) Hsu, C.X. Zhai and K. 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, **(2013)**.

[13] K. Chakrabarti, S. Chaudhuri, T. Cheng and D. 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, **(2012)**.

[14] M. Harada, S.-Y. Sato and K. Kazama, "Finding Authoritative People from the Web", In Proceedings of the Joint ACM/IEEE Conference on Digital Libraries (JCDL04), **(2004)**, pp. 306–313.

[15] K. Dmitri, V. Z. (Stella) Chen, S. Mehrotra and R. Nuray-Turan, "Web People Search via Connection Analysis", IEEE Transactions on Knowledge and Data Engineering, vol. 20, no. 11, **(2008)**, pp. 1–16.

[16] D. Bollegala, Y. Matsuo and M. Ishizuka, "Automatic Discovery of Personal Name Aliases from the Web", IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 6, **(2011)**, pp. 831–844.

[17] X. Ren and T. 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, **(2015)**.

[18] R. Blanco, B. Barla Cambazoglu1, P. Mika and N. Torzec, "Entity Recommendations in Web Search", In International Semantic Web Conference, **(2013)**.

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

Authors

Mamta Kathuria, she received her MCA degree from Kurukshetra University and M.Tech from MDU, Rohtak in 2008, respectively.. She is pursuing her Ph.D in Computer Engineering from YMCA University of Science and Technology, Faridabad. She is currently working as a Assistant Professor in YMCA University of Science & Technology and has eight years of experience. Her areas of interest are search engines, Web Mining and Fuzzy Logic.

Anuradha, she received B. Tech in Computer Engineering from YMCA university of Science & Engineering, Faridabad in 2014. She is pursuing her M. Tech in Computer Engineering from YMCA University of Science and Technology, Faridabad. Her areas of interest are Search Engines, Information Retrieval and Enhancing the quality of Web Search.

Chander K. Nagpal, he is Ph. D (CS) from Jamia Milla Islamia, New Delhi. He is currently working as a professor in YMCA University of Science & Technology and has twenty eight years of teaching experience. He has published two books. He has published many research papers in reputed international Journals such as IEEE transaction, Wiley STVR, CSI. His academic interests include Ad hoc networks, Web Mining and Soft Computing.

Neelam Duhan, she received M. Tech in Computer Engineering from MDU, Rohtak in 2005. She completed her PhD in (CE) in 2011 from Maharshi Dayanand University, Rohtak. She is currently working as an Assistant Professor in CE Department in YMCA University of Science and Technology, Faridabad and has an experience of about 12 years. She has published over 30 research papers in reputed international Journals and Conferences. Her areas of interest are databases, search engines and web mining.