Interactive Query Expansion using Concept-Based Directions Finder Based on Wikipedia

Yuvarani Meiyappan[1] and Sriman Narayana Iyengar[2]
[1]Education and Research at Infosys limited, India
[2]School of Computing Science and Engineering, VIT University, India

Abstract: *Despite the advances in information retrieval the search engines still result in imprecise or poor results, mainly due to the quality of the query being submitted. The query formulation to express their information need has always been challenging for the users. In this paper, we have proposed an interactive query expansion methodology using Concept-Based Directions Finder (CBDF). The approach determines the directions in which the search can be continued by the user using Explicit Semantic Analysis (ESA) for a given query. The CBDF identifies the relevant terms with a corresponding label for each of the directions found, based on the content and link structure of Wikipedia. The relevant terms identified along with its label are suggested to the user for query expansion through the new visual interface proposed. The visual interface named as terms mapper, accepts the query, and displays the potential directions and a group of relevant terms along with the label for the direction chosen by the user. We evaluated the results of the proposed approach and the visual interfacefor the identified queries. The experimental result shows that the approach produces a good Mean Average Precision (MAP) for the queries chosen.*

1. Introduction

Web search is one of the most predominant means for acquiring information on any topic of interest. Finding the relevant and useful information is more challenging due to the increase in the volume of information available and not being organized properly. Research in this field includes improving the relevance of the search results for the given query and formulating the query for better precision (formally termed as query expansion). Query expansion is the topic of interest in this paper.

The formulation of query by users has always been a challenge as the web is being used by a diverse population varying in their level of expertise. The users of the web search engines range from novices to experts in specific topics being searched. The experts who are familiar with the topic could formulate appropriate query for better search, a user. however, often has only a vague idea of the relevant terms and relies on an iterative process to determine the terms from the retrieved query results.

It has been found that most users are not aware of the search process, in a hurry or in a habit of specifying short queries with little context. Most queries used by larger public are incomplete, inappropriate or not precise [6]. The uncommon, nonspecific and ambiguous queries would result in non-relevant results. Even in B2B ecommerce applications, the keyword based search results in limited information extraction [9]. The study [12] suggests that only approximately 10% of the search is navigational and transactional, more than 80% of the searches are being informational in nature. The better the quality of the queries, the better the quality of the search results generated [3]. Many methods have been devised to help the user in modifying the query to produce better precision, including the automatic and interactive query expansion. The automatic query expansion system adds the identified terms to the query by itself. The interactive query expansion suggests the expansion terms to the user to indicate the term of choice. The process of suggesting the terms for expansion is referred as term suggestion.

The global methods to achieve query expansion includes the query reformulation with a thesaurus or WordNet, query expansion through automatic thesaurus generation, techniques like spelling correction, query log mining [2]. The query expansion has also, been achieved through the label of the search result clusters. Some of the most popular search engines like Yahoo, Google suggest the related term for the query submitted along with the search results. Google suggests the terms as the user types them in the user interface. To the best of our knowledge, the term suggested even by most of the popular search engines are not up to satisfaction. For instance, even the popular search engines do not suggest Jaguar XF, XJ and XK car models for "Jaguar" [11, 26].

This paper proposes a general-purpose interactive query expansion methodology (named CBDF). It

allows the user to choose the interested topic interactively. The challenge for query expansion is that the same idea can be expressed by various terms or phrases that could be a variation, abbreviation, acronyms, nicknames, aliases and others [10]. The objective of the proposed methodology is to suggest the terms that are potentiallyrelated to the given query and leads the searcher to the target topic and hence the search results, quickly. This has been achieved through identifying and suggesting the possible directions and presenting the grouped terms related to directions of user's interest. Initially the Concept-Based Direction Finder (CBDF) identifies the potential direction for the user using the Explicit Semantic Analysis (ESA) proposed by Gabrilovich in [8]. This would enable the user to move further in a directionintended. A component termed CBTermExtractor has been implemented to suggest the grouped relevant terms to the user for the direction preferred. Borrowing the concept of link structure and its link text in general web, the proposed CBTermExtractor utilizes the link structure of Wikipedia in identifying the related terms for term suggestion and the structure of Wikipedia content in grouping them, enabling the query expansion. A new visual interface has been proposed to present the directions, label and the grouped terms to the user, as shown in Figure 1.
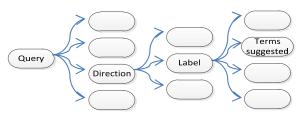

Figure 1. Directions and terms suggestion.

Our contributions in this paper are two manifolds. First, the relevant terms from Wikipedia are being identified based on the bidirectional link between the identified source Wikipedia article and the Wikilinks within it. To the best of our knowledge, utilizing the bidirectional link between the Wikipedia articles for identifying the semantic relation between the terms is novel. Second, a visual interface is being proposed that presents the direction, concept under which the suggested terms are grouped and the strong and semantically related terms, as shown in Figure 1.

The rest of this article is organized as follows: The related works are presented in section 2. The Wikipedia link structure, being the backbone of this proposed approach is explained in section 3. An overview of ESA has been discussed in section 4. CBTermExtractor, the component that identifies the relevant terms is discussed in section 5. In section 6, the overview of the CBDF has been proposed. Followed by CBDF, its experimental results are produced in section 7. Finally, Section 8 has the conclusion with the further work.

2. Related Work

The approaches being used for interactive query expansion predominantly can be classified into concept-based or based on clustering the search result. The work we have proposed here is concept-based and hence the existing works that are related to concept-based alone are briefed here, excluding the work related to clustering the search results.

The concept-based approaches identifies the relationship between terms based on manually or automatically constructed thesauri, ontologies, term co-occurrences, latent word relationships and concepts from a corpus [5]. The domain specific concept-bases are ones built by the coherent community of experts. They. However, cannot be used for general-purpose. The manual generation of domain-specific lexicon is expensive as it involves domain experts. With the rapidly emerging needs, it is difficult to cover the new or domain-specific terminologies. There are general-purpose dictionaries and thesauri such as WordNet, which can be used for general-purpose query expansion. These approaches. However, have poor coverage of proper nouns related to the given query. In this proposed work, the Wikipedia which is a web-based collaborative encyclopedia has been used as the concept-base.

As explained by Syafrullah and Salim [18], the term extraction tasks are generally based on linguistic methodologies, terminology and Natural Language Processing (NLP) method, or approaches based on statistical/information retrieval methods. The linguistic method for term extraction follows shallow text processing techniques that include tokenizer, part-of-speech tagger, and syntactic analyzer. The terminology and NLP methods perform the internal analysis within the corpus. The statistical methods are based on the comparison of frequencies between domain-specific and general corpora, unlike the terminology and NLP approach. The approach being proposed here is based on terminology and NLP.

In [14], Milne proposed an approach that identifies the corresponding Wikipedia article for the given query and the Wikipedia Link Vector Model (WLVM) technique has been used to determine the relationship between the terms. WLVM technique measures the semantic similarity based on the angle between the vectors of the links found in each corresponding Wikipedia articles. The vectors are generated based on weight of each link within the article, instead of the term count (tf-idf). The weight has been measured based on the number of times the article is linked to a target article with inverse probability of any link pointing to the target article.

The approach proposed in [10], addresses the problem of variable terminology by constructing the term-concept map based on Wikipedia article. The articles that has bidirectional link are considered for

concept-to-concept relationship. The approach however considers only the title of the Wikipedia articles as concepts, ignoring the content of the page.

Li *et al.* [13], proposed an approach to expand the ad-hoc queries using Wikipedia. The approach identifies the related Wikipedia articles, extracts the categorical information of these articles and groups them based on the category. It then identifies the top 40 terms from 20 articles to expand the query.

Mima *et al.* [16], has proposed Automatic Term Recognition (ATR) system for the biomedical domain to identify the pertinent terms and their variants from a document collection. The C/NC-value automatic term-recognition method [15], has been used to help domain experts to gather and manage domain-specific terminologies. The method enhances the common approaches (frequency of occurrence) by considering the type of terms which includes nester terms and multiword terms.

Velardi *et al.* [21], extracted the terms using NLP methodologies. The terminology extraction process accepts the domain-relevant documents and applies the part-of-speech tagging and chunking on the input document. The extracted candidate terms are filtered based on domain relevance, domain consensus, lexical cohesion, structural relevance and heuristics. The user manually validates the page for the terminologies extracted. Apart from being manually validated for terms, the accuracy of the results of this approach highly depends on the domain-relevant document being provided.

The domain-specific lexicon is of utmost importance in NLP and information access. Avancini *et al.* [1], proposed an approach for the automatic expansion of domain-specific lexicon from initial lexicon into a larger lexicon. The authors implemented text categorization task (referred as term categorization) with terms represented as vectors in the space of documents. The AdaBoost.MH^{KR} classifier algorithm has been used to classify the terms. The system however can be improved further as the F1 values are still low.

Syafrullah and Salim [18], has proposed a new approach using particle swarm optimization techniques to extract the terms. The approach considers the domain relevance, domain consensus, term cohesion, first occurrence and length of noun phrase, for features in order to extract the terms. A weight is assigned to the term based on the five features and the particle swarm optimization is applied to determine the appropriate set of feature weights.

3. Wikipedia Link Structure

Wikipedia articles that are related to each other are well linked or cross-referenced [22]. The highlighted text within the Wikipedia article refers to the relevant Wikipedia article with further in-depth information about the text (referred as Wikilinks [24]). Wikipedia article includes other links to articles of interest, relevant external websites and pages included in the citation (referred as external links), reference material, links to pages on another Wikimedia project website (referred as Interwiki links), links to a section within the page (referred as section linking), ISBN links and organized categories of knowledge. With various types of links available within Wikipedia, this research focuses on only Wikilinks.

3.1. Wikilinks

As per the style for Wikipedia articles manual [23], the general guideline for "what should be linked" within Wikipedia with respect to Wikilinks states that links should be created to relevant article that will help readers understand the article more fully. The relevant article includes the article about people, event, technical terms, jargon, slang expression and others. The guideline for "what should not be linked" states that links should be avoided for the plain English words, major geographic features and locations, religions, languages, common profession, common units of measurement, dates and others.

3.2. Anchor Text

The anchor text is the descriptive labeling of the hyperlink defined within a page which is visible to the viewer [25]. Wikipedia guideline states that the anchor text linked to an article should correspond to the Wikipedia article as close as possible, given the context. In general, the anchor text is usually believed to be relevant to the page being referred.

3.3. Forward / Backward Link

Wikipedia articles, like any web page has the forward and backward links. The backward links are the incoming links to a web site or page. here in this work, we refer to the backward links as incoming links to the Wikipedia article from another Wikipedia article. The forward links are the outgoing hyperlinks from a web site or page. here we refer the outgoing links to the other Wikipedia article as forward links. The source page that has the Wikilink (forward link) is termed as anchor page. The page which the Wikilink in the source page refers to is termed as target page.
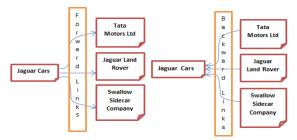

Figure 2. Forward and backward links.

In Figure 2, the anchor text Tata Motors Ltd is a Wikilink as it refers to another Wikipedia article titled "Tata Motors Ltd". The Wikipedia article titled "Jaguar Cars" is the source page for the Wikilink "Tata Motors Ltd" and the Wikipedia article titled "Tata Motors Ltd" is the target page. The figure demonstrates the forward and backward link in it.

When there is more to explain about a term in a Wikipedia article, it will be linked to the relevant article within Wikipedia. For instance, the terms like Coventry, Tata Motors Ltd, Jaguar Land Rover, Swallow Sidecar Company, Sir William Lyons, sidecars has Wikilinksin the Wikipedia article titled "Jaguar Cars". These terms have further in-depth information explained in their corresponding target article. In Wikipedia, the links between relevant Wikipedia articles are strong. The CBTermExtractor explained in section 5, considers this as the core for identifying the relevant terms for the given Wikipedia concept.

4. Explicit Semantic Analysis: An Overview

Traditionally the semantic features of documents and queries were represented either by constructing taxonomy of concepts and relations manually or automatically or by analyzing the latent relationship between terms within the document or query [4]. In the former approach, the system should map the given text or document to a set of nodes within the taxonomy restricting to fewer concepts, whereas the later approach might result in concepts the user might not be aware of Gabrilovich and Markovitch. however, in their work [7] proposed the third approach termed as ESA.

ESA initially identifies each Wikipedia article as concept, generates the inverted index with attributes belonging to various concepts (mapping to the corresponding concept that contains the attribute), classifies the text onto the Wikipedia concepts using centroid classifier-representing each concept with an attribute vector of the article text. The classifier ranks the Wikipedia concepts by their relevance to the given text. The ESA approach generates the feature using a multi-resolution approach. The features are first generated at the individual word level followed by sentences, paragraphs before the entire document. The feature selector eliminates the extraneous features among the generated. ESA has already been proved to be effective in text categorization [7], computing semantic similarity [8].

5. CBTermExtractor: Term Extractor for a Given Wikipedia Concept

CBTermExtractor is a term extractor specific to term suggestion for a given topic. The lexicon generated by CBTermExtractor should contain terms that suggest

search topics for the given concept. Typical term extractors parses the text, extracts terminological structures and purge out the terms based on filters. The term extractors would filter for terms with high frequency either within a document or across the corpora, terms that occur together frequently, structural terms that are highlighted within the document or corpora. These filters would result in a large lexicon and each term might not add value always while expanding the query. Figure 3 shows the extract from Wikipedia for "Jaguar Cars" with the nouns being highlighted. Stanford Log-linear Part-Of-Speech tagger [20], has been used to identify the possible nouns in the extract. Not all the highlighted terms when expanded with the query term may result in better precision or recall.

Jaguar Cars Ltd, known simply as *Jaguar*(◀) /ˈdʒæɡjuːˌɑr/) is a British luxury car manufacturer, headquartered in Whitley,Coventry, England. A wholly owned subsidiary of the Indian company Tata Motors Ltd., it is operated as part of the Jaguar Land Rover business. [2]

Figure 3. Nouns in Jaguar cars (extraction from Wikipedia).

CBTermExtractor focuses on extracting those terms that suggest semantically relevant terms, representing a concept within the given topic during information retrieval. The relatedness between the terms and the topic should be strong that the term could be further explored by expanding the query. The component considers the Wikipedia link structure to build the lexicon for the given topic.

CBTermExtractor considers these Wikilinks between the relevant articles to determine the relevant terms for the domain-specific lexicon. Figure 4 shows the link structure of Wikipedia articles. Each node represents the Wikipedia article, straight lines from the node represent the forward link to the target Wikipedia article, dotted link represents the backward link from the target page (for the forward link) to the source Wikipedia article. The Wikilink becomes strongly relevant and related to the concept, if the source article has both forward link(s) to as well backward link(s) from the target.
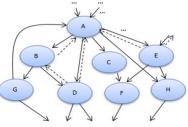

Figure 4. Graph structure of Wikipedia.

For instance, as shown in Figure 2, the Wikilink with anchor text Tata Motors Ltd has backward link from the target to the current page which is the Jaguar cars (source Wikipedia page). This demonstrates that there is in-depth relatedness between the source and

the target page. The CBTermExtractor considers these Wikilinks to be strongly and semantically related to the given Wikipedia concept.

The CBTermExtractor considers the Wikilink to be more relevant if there is both backward as well forward links between the source and target articles. It considers each Wikilinks in the source page which has forward as well backward link to be more related to the given topic. In Figure 2, for the topic "Jaguar Cars", the "Tata Motors Ltd" is a Wikilink which has forward link to the Wikipedia article titled "Tata Motors Ltd". The Wikipedia article titled "Tata Motors Ltd" has a backward link to the source page (Jaguar Cars) of the Wikilink as well. The anchor text of these Wikilinks would be considered to be the candidate term for the topic suggestion and hence added to the concept-specific lexicon. On the other hand, the Wikilink with anchor text "London Stock Exchange" does not have a backward link from Wikipedia article titled "London Stock Exchange" to "Jaguar Cars" article. Hence the term "London Stock Exchange" is not considered to be more relevant to be included as a suggested term for "Jaguar car" and hence not included in the concept-specific lexicon. For the given query q, $WS(i)$ is a specific direction. $WS(i)$ is represented as:

$$WS(i)=\{WS(i,1),\ WS(i,2),\ ...,\ WS(i,n)\} \qquad (1)$$

where $WL(i,n)$ is the Wikilink from $WS(i)$. The $WS(i)$ denotes the Wikipedia source article (direction) for which the related terms ($Terms(i)$) are to be identified. Let $WTT(i,x)$ represents the xth anchor text within $WS(i)$ Wikipedia source article which links to $WT(i,x)$ Wikipedia target article, bidirectional. The concept-specific lexicon for the concept, $Terms(i)$ for $WS(i)$ Wikipedia source article is represented as:

$$Terms(i) = \left\{ WTT(i,x) \left| \begin{array}{l} \exists\ wikilink \in WT(i,x): \\ wikilink\ is\ backward \\ link\ to\ WS(i) \end{array} \right. \right. \qquad (2)$$

where, $WS(i)$ is the source Wikipedia article of ith direction; $WE(i,x)$ is the target Wikipedia article, the xth Wikilink from the source Wikipedia article ($WS(i)$) to the target Wikipedia article $WTT(i,x)$ is the title of the target Wikipedia article ($WT(i,x)$).

6. Concept-Based Directions Finder (CBDF)

The CBDF system identifies the directions and the potential terms of search in the direction chosen to be suggested to the user. Figure 5 shows the overview of the CBDF system.

6.1. Overview

The CBDF system has 3 components, namely Direction Identifier, Term Suggestor, and Terms Mapper.
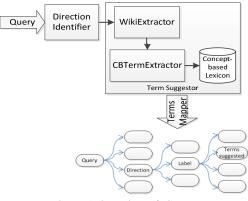

Figure 5. Overview of CBDF.

6.1.1. Direction Identifier: Identifies the Directions Based on ESA

The direction identifier is the component that accepts the query from the user and identifies the possible directions for the given query. The component accepts the user query and invokes the ESA for identifying the directions. The Research-ESA is an open source implementation of ESA [17], which is being used for ESA results. The query is submitted to the Research-ESA Explicit Semantic Analysis which is exposed as a Web Service. The result is a list of terms that represent the meaning of the given term in a high-dimensional space of Wikipedia concepts. The component considers these concepts as directions for the user to continue search.

6.1.2. Term Suggestor: Identifies the Strongly Related Terms for the Direction

Term Suggestor is the component of CBDF that accepts the directions from the Direction Identifier component. The component identifies the terms that has semantic and strong relationship with the direction.

The component accepts the list of directions from the direction identifier component. The component has WikiExtractor and CBTermExtractor. The WikiExtractor extracts the corresponding Wikipedia article for each of the directions from the direction identifier. It is implemented as Java thread class for better performance. The extracted Wikipedia content is provided to the CBTermExtractor.

The CBTermExtractor identifies the terms that are strongly related to the direction and stores them in the concept-based lexicon, as explained in section 5. The component accepts the source Wikipedia article from WikiExtractor. It parses through the article for Wikilinks within. The CBTermExtractor extracts the content of these Wikilinks with the help of WikiExtractor. The CBTermExtractor generates the Terms$_i$ for the accepted source Wikipedia article. The component generates the Terms(i) for the entire source Wikipedia articles identified as directions by the direction identifier component. The Terms$_i$ for each direction are stored in the concept-based lexicon.

The CBTermExtractor of CBDF groups the terms identified into specific concepts. For instance, the Jaguar XF, Jaguar XJ and Jaguar XK are car models, whereas the Jaguar AJ-V8, Jaguar AJ-V6 and Jaguar AJD-V6 are the engines designed for Jaguar cars. The CBTermExtractor groups these separately and label them. The component considers the title under which these anchor text occurs for grouping and labeling them. The label for the groups is the corresponding Wikipedia sub-title under which they occurred. The terms are stored in the concept-based lexicon repository associated with the group label identified.

6.1.3. Terms Mapper: Presents the Directions and Terms to the User

Terms Mapper component presents the directions, terms and their mappings stored in the concept-based lexicon to the user. The user interface named Terms Mapper is novel, up to the best of our knowledge. This is the first of its kind interface which accepts the query from the user and displays the potential directions for the user to continue the search, the concepts within the direction and the suggested terms. The interface accepts the query from the user and presents the potential directions for the accepted query. The user, upon selecting the direction to be explored further, is presented with the list of concepts within the direction to be explored further. Once the user opts for the concepts, the interface displays the relevant terms to the user.

Figure 6 shows the terms mapper. The left node represents the user's query. The immediate nodes on the right to the query represent the potential directions for the user's query. The immediate nodes on the right to the directions node represent the concepts under which the strong and semantic related terms are grouped. The right nodes (next to the concepts) are the suggested terms from the CBDF.
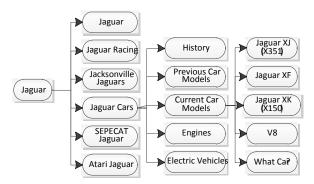

Figure 6. Terms Mapper for 'Jaguar' keyword.

7. Experiments

7.1. Background Information

The various evaluation protocols used for evaluating the term extractor includes the Gold standard that has the standard set of terms for a given domain,

comparing the result of the proposed term extractor with the existing ones or based on survey.

Terms can be related to each other as synonymy, hyponymy, meronymy, holonymy, troponymy, hypernym, polysemy and others. The proposed system extracts the terms as relevant, if the terms are related to the query in the corresponding context without considering the relationship between them. Most importantly, the system accepts only the query as input and does not accept the document collection from which the terms have to be extracted, unlike most of the term extractors proposed. Typically the objective of the system is to suggest the direction and related terms to the users ranging from novices to experts.

Determining the relevance of the term to the given topic is challenging here, as the user might be interested in different information. For instance, a user who is searching about cricket might be interested in exploring about a particular bowling technique (googly, Yorker), the players, the powerplay, fielding positions or others.

To the best of our knowledge, there is no data set available to evaluate the proposed term extractor which suggests the direction of search, grouped relevant terms with label together. Hence we decided to evaluate the proposed approach through survey.

The Wikipedia article for each of the identified concepts are extracted from the Wikipedia dumps downloaded from http://download.wikipedia.org. A multithreaded Java program has been implemented to extract the Wikipedia articles. The Wikilinks within the articles are identified. The program then identifies whether there is a backward link from the target Wikipedia article.

7.2. Evaluation Metrics

The proposed CBDF has been evaluated for Precision (P) and Mean Average Precision (MAP). Both the precision of the suggested terms for the preferred direction (P(Direction)) and the precision of the suggested terms for the labels (P(Label)) have been measured. The Precision (P) is measured as:

$$P = \frac{\left|\{relevant\ terms\}\right| \cap \left|\{retrieved\ terms\}\right|}{\left|\{retrieved\ terms\}\right|} \quad (3)$$

The MAP measures the average precision for each query by determining the mean of the precision. Similar to the precision, the MAP has also, been measuredfor both the direction as well the label.

7.3. Experiments and Results

Including the queries from the TREC-Web track [19], we identified 20 queries for the evaluation. We requested 25 graduate and post graduates to evaluatethe results. They were given the results of the 20 queries. The first 10 directions for each query is

considered for evaluation. Each direction is mapped to an average of 7 labels. Each label is mapped to an average of 14 suggested terms, excluding the terms suggested under 'general'. The participants were requested to rate the relevance of the suggested terms to the directions and the label under which they are grouped. The participants rated the suggested terms as relevant, irrelevant or not sure. Based on the survey, the average precision for each query has been calculated based on the survey response from all the evaluators.

Figure 7 shows both the precision for the suggested terms based on both the direction as well the label. The MAP based on the direction is 0.76 and the MAP based on the label is 0.58.
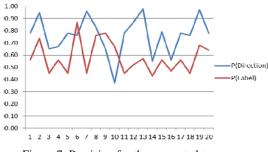

Figure 7. Precision for the suggested terms.

Along with the evaluation of the relevance of the suggested terms, the evaluation of the visual interface was also, performed. The evaluators were requested to rate the usefulness of the visual interface in the scale of 1 to 5, 1 being rated low and 5 being high. The user interface was given 4.6 out of 5.

8. Conclusions

We presented a novel approach for interactive query expansion based on Wikipedia, introducing a feature selection component, CBTermExtractor. The result shows that the approach identifies the strong and semantic related terms to the given query along with the different directions. The work however can be extended further to enhance the ESA for identifying the directions, as the document representation of ESA can be optimized. The precision of the suggested terms for the labels (P(Label)) also, suggests that it can be further improved.

References

[1] Avancini H., Lavelli A., Sebastiani F., and Zanoli R., "Automatic Expansion of Domain-Specific Lexicons by Term Categorization," *ACM Transactions on Speech and Language Processing*, vol. 3, no. 1, pp. 1-30, 2006.

[2] Christopher M., Prabhakar R., and Hinrich S., *An Introduction to Information Retrieval*, Cambridge University Press, Cambridge, 2008.

[3] Croft B. and Thompson H., "I3R: A New Approach to the Design of Document Retrieval Systems," *Journal of the American Society for Information Science*, vol. 38, no. 6, pp. 389-404, 1987.

[4] Egozi O., Gabrilovich E., and Markovitch S., "Concept-Based Feature Generation and Selection for Information Retrieval," *in Proceedings of the 23rd National Conference on Artificial Intelligence*, Chicago, vol. 2, pp. 1132-1137, 2008.

[5] Egozi O., Markovitch S., and Gabrilovich E., "Concept Based Information Retrieval using Explicity Semantic Analysis," *ACM Transactions on Information Systems*, vol. 29, no. 2, pp. 1-34, 2011.

[6] Fonseca B., Golgher P., Pôssas B., Ribeiro-Neto B., and Ziviani N., "Concept-Based Interactive Query Expansion," *in Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, Germany, pp. 696-703, 2005.

[7] Gabrilovich E. and Markovitch S., "Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge," *in Proceedings of the 21st National Conference on Artificial Intelligence*, vol. 2, pp. 1301-1306, 2006.

[8] Gabrilovich E. and Markovitch S., "Computing Semantic Relatedness using Wikipedia-Based Explicit Semantic Analysis," *in Proceedings of the 20th International Joint Conference on Artificial Intelligence*, USA, pp. 1606-1611, 2007.

[9] Ghobadi A. and Rahgozar M., "An Ontology-Based Semantic Extraction Approach for B2C e-Commerce," *International Arab Journal of Information Technology*, vol. 8, no. 2, pp. 163-170, 2011.

[10] Gregorowics A. and Mark K., "Mining a Large-Scale Term-Concept Network from Wikipeida," *Technical Report*, MITRE Corporation, USA, 2006.

[11] Google, available at: http://www.google.com, last visited 2011.

[12] Jansen B., Booth D., and Spink A., "Determining the Informational, Navigational, and Transactional Intent of Web Queries," *International Journal on Information Processing & Management*, vol. 44, no. 3, pp. 1251-1266, 2008.

[13] Li Y., Luk P., Ho S., and Chung F., "Improving Weak Ad-hoc Queries using Wikipedia as External Corpus," *in Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Amsterdam, pp. 797-798, 2007.

[14] Milne D., "Computing Semantic Relatedness using Wikipedia Link Structure," *in Proceedings of the New Zealand Computer Science Research Student Conference*, New Zealand, pp. 1-8, 2007.

[15] Mima H. and Ananiadou S., "An Application and Evaluation of the C/NC-Value Approach for the Automatic Term Recognition of Multi-Word Units in Japanese," *International Journal Terminology*, vol. 8, no. 2, pp. 175-194, 2001.

[16] Mima H., Ananiadou S., and Matsushima K., "Terminology-Based Knowledge Mining for New Knowledge Discovery," *ACM Transactions on Asian Language Information Processing*, vol. 5, no. 1, pp. 74-88, 2006.

[17] Research-ESA Web Service, available at: http://www.multipla-project.org/research_esa_ui/configurator/index/, last visited 2011.

[18] Syafrullah M. and Salim N., "Improving Term Extraction using Particle Swarm Optimization Techniques," *Journal of Computing*, vol. 2, no. 2, pp. 116-120, 2010.

[19] TREC 2010 Web Track, available at: http://trec.nist.gov/data/web10.html, last visited 2011.

[20] Toutanova K., Klein D., Manning C., and Singer Y., "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network," *in Proceedings of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, USA, pp. 252-259, 2003.

[21] Velardi P., Navigli R., and D'Amadio P., "Mining the Web to Create Specialized Glossaries," *IEEE Intelligent Systems*, vol. 23, no. 5, pp. 18-25, 2008.

[22] Wikipedia, available at: http://en.wikipedia.org/wiki/Wikipedia:About#Basic_navigation_in_Wikipedia, last visited 2011.

[23] Wikipedia-Manual, available at: http://en.wikipedia.org/wiki/Wikipedia:Linking, last visited 2011.

[24] Wikipedia-Help: Link, available at: http://en.wikipedia.org/wiki/Help:Link, last visited 2011.

[25] Wikipedia, available at: http://en.wikipedia.org/wiki/Backlink, last visited 2011.

[26] Yahoo, available at: http://www.yhoo.com, last visited 2011.

Sriman Narayana Iyengar obtained his MSc, ME, PhD. Currently, he is director for Perivar EVR Central Library and senior professor at the School of Computing Science and Engineering at Vellore, India. His research interests include agent based distributed computing, security aspects of all networks including VOIP, intelligent information retrieval, computational methods, bio informatics and fluid mechanics. He has authored and co-authored several books and had nearly 120 research publications in reputed peer reviewed international Journals. He served as PCM/Reviewer for many international and IEEE conferences. He is a chief editor for IJSEA of AIRCC, guest editor for special issue on "cloud computing and services" of Int'l J. of communications, network and system sciences. He is also, an editorial board member for many reputed international journals like IJCA, IJCTE, IJSE, IJEMTA, JCMS, and many more.

Yuvarani Meiyappan is working as Lead in Education and Research at Infosys limited, India. Currently, she is doing her PhD in VIT University. Her research interest includes information retrieval, machine learning and semantics.