Information Retrieval Based on Extraction of Domain Specific Significant Keywords and Other Relevant Phrases from a Conceptual Semantic Network Structure

Mohammad Moinul Hoque
University of Evora, Portugal
moincse@yahoo.com

Prakash Poudyal
University of Evora
prakashpoudyal@gmail.com

Teresa Goncalves
University of Evora,
tcg@uevora.pt

Paulo Quaresma
University of Evora
pq@uevora.pt

ABSTRACT

This paper presents a functional approach towards the problem domain of Information Retrieval System built upon a narration based search text. The presented system retrieves documents from the background collection by extracting the domain specific significant keywords and other relevant phrases from a given narrative search text. The narrative search text can be a description or scenario which poses a great difficulty in the problem domain to retrieve the relevant document sets with efficiency and accuracy from the background data repository. We have adopted a different approach where the significant keywords are extracted from the narration text to form a search query and alternative sets of queries are also formulated by expanding the search query from a Conceptual Semantic Network built for the purpose. Inclusion of the Conceptual Semantic Network and WordNet synonym sets for the search query expansion plays an important role in the retrieval mechanism. Experiments were carried out on the data sets from the Ad-hoc task of 'Forum for Information Retrieval Evaluation, 2013'. The background data set contained a huge number of legal documents consisting of data over 3 GB and was divided into two domains such as 'Consumer Law' and 'Hindu Marriage & Divorce Law'. For the search queries, a set of scenarios in the form of narrative text were provided. The system was required to perform an analysis of the search text and retrieve a set of top 1000 legal documents for each of the queries from the background collection which may be relevant to the situation described in the narration of the search text.

I. INTRODUCTION

Information retrieval is generally explained as an activity of obtaining relevant information from the vast collection of resources lying in background [1]. Searches are usually made on the background data repository based on the search text [2]. Searching can be significantly expensive when the data source is huge in size and retrieving the most relevant documents asks for some well defined mechanism so that the retrieval accuracy can be increased. Moreover, defining a good structure or representing the background information in some suitable form and accessing them in quick time are also a matter of concern. To a certain extent, constructing a logical structure or index with the background data and searching in the structure requires a formal mechanism to ensure the retrieval efficiency. Automated indexing [3] that are used in the present form of Information Retrieval (IR) systems encounters a few problems that can be presented in terms of nature of information stored and of course the kind of users expected to use the IR systems using in different manner. A good IR system should have the ability to take a user query and return as many relevant documents as possible while rejecting the non-relevant ones. Selection of appropriate terms or keywords from the

search text is also a major area of concern as it directly impacts on the retrieval accuracy. Having a good precision and high recall [4] are the two common aspects used in IR as a measure of how good the system is.

2. PROPOSED SYSTEM

In this work, we are dealing with domain specific corpus for legal documents and the search space is also domain dependent. This allows us to discover a potential model in which we propose to build a Concept based Semantic Network structure (CSN) manually. CSN contains various conceptual terms/phrases related to the respective domains and connection among them. CSN is consulted when generating the search queries from the narrative search text. For example, in the Hindu Marriage Law domain, the following figure (1.a) can be seen as a partial view of the domain specific conceptual semantic network. A simple and partial conceptual network for the consumer law based domain is depicted in the figure 1(b). As shown in the figure 1(a) and in 1(b), the directed arrows show the dependency of the concepts with the other concepts. These concepts are extracted from Wikipedia using a crawler application developed for the purpose. We found the conceptual network to be very useful when searching for various extractable keywords from the search text for generating one or more search queries. Significant keywords extracted from the search text and combined with Named Entities are used for the document retrieval process which is explained in the subsequent sections.

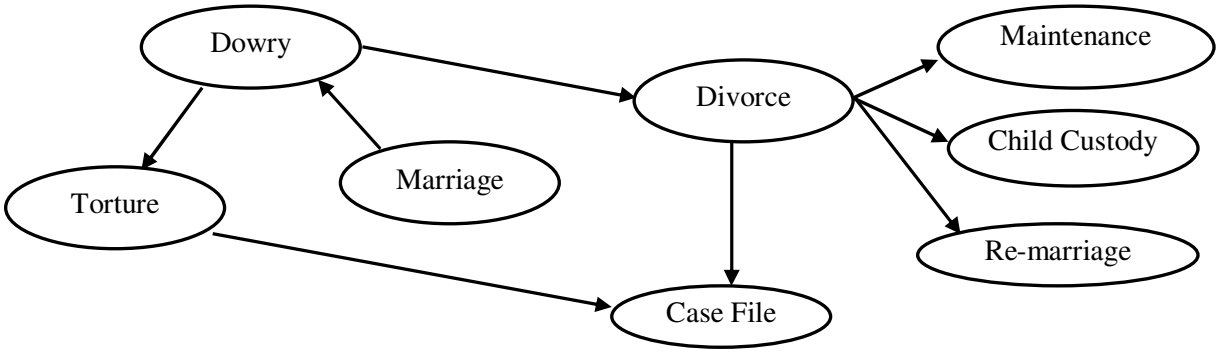

Figure 1(a): A partial view of the domain specific (Hindu Marriage and Divorce Law) Conceptual Semantic Network

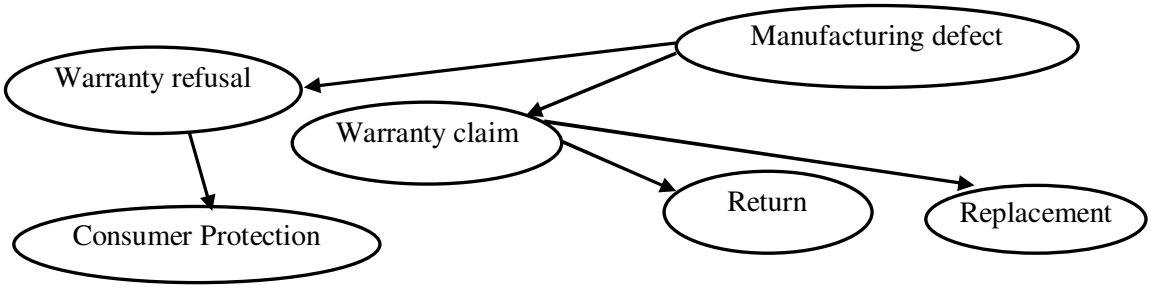

Figure 1(b): A partial view of the domain specific (Consumer Protection Law) Conceptual Semantic Network

2.1. Indexing the document corpus

The first module in our system indexes the background data. Before getting into the indexing step, the background data is preprocessed first by stripping off a few data structure tags and then stemming is performed using the Porter Stemming Algorithm [5] over the data. Finally the data set gets indexed using an inverted index [6] structure. These steps are shown in figure 2 below.

The background collection contains the verdicts from the Supreme Court and various acts of parliament related to the Consumer Law domain and the Hindu Marriage & Divorce Law domain and depending on the background collection data of FIRE 2013 [7], three separate indices were created, one for each of the Consumer Law data and Hindu Marriage & Divorce Law data and a combined index containing both sets of data.

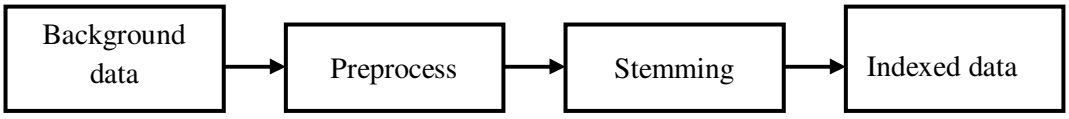

Figure 2. Creation of a structured inverted index from the background collection

2.2 Search text analysis and processing

Instead of simple keywords or a combination of keywords leading to a sentence, the search queries that we analyzed from the FIRE 2013 query set are descriptive narration that tells the story of a situation. We are supposed to retrieve the documents that are relevant to the description. Before getting into the actual searching space and ranking of the retrieved documents, we convert the narrative description into a structured search text by adopting a few relatively different approaches.

2.2.1 Preprocessing

First of all, the given narration is preprocessed in which the English stop words are eliminated from the text since they are very less or not significant at all in the case of retrieving documents and is followed by a step where the text is freed from the noisy symbols or characters. After these two steps, the search text is converted into a set of sentences using the heuristics method employed by the OpenNLP's API [8].

2.2.2 Parts of Speech tagging

In this step, the words in the preprocessed search sentences are tagged with their corresponding Parts of Speech (POS) tag. These tags will be useful for building alternative search queries with possible synonym sets of the keywords and will be explained in the next sub-sections. We have used the Stanford POS tagger [9,10] for tagging the POS labels for possibly significant keywords.

2.2.3 Marking the Named Entities

Named-Entity (NE) recognition is a subtask of information extraction that identifies, locates and classifies various elements in the corpus into predefined categories such as the names of persons, names of organizations, locations, expressions of times, quantities, monetary values etc [11]. In this step the NEs are recognized from the search sentences and are stored for future use. For the NEs detection, we were specifically concerned about the identification of the organization and locations only.

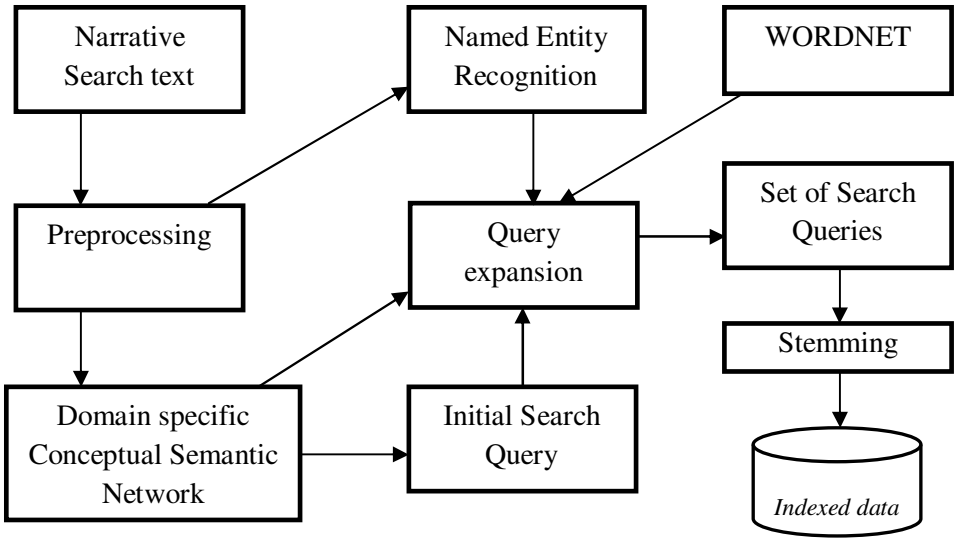

Figure 3. Search text analysis, processing and query expansion

These entities will be used for expanding the query text. From the preprocessed search text, the non-stop words are initially picked as possible keywords to build up an initial search query.

2.3 Search Query generation using a Conceptual Semantic Network and synonym set.

Once, we have the initial set of keywords to form a search query, our query expansion module searches the keyword set and marks the domain specific significant keywords using the nodes of the Conceptual Semantic Network (CSN). Based on the parts of speech tags of the marked keywords, a possible set of synonyms are also extracted using WORDNET synset [12]. These synonyms are also added to create an alternate set of queries for possible file retrieval performance enhancement. The steps performed for search query generation and search query expansion are shown in figure 3.

Let's take a look at a portion of an example search narration S1: *"I am a Hindu girl married for over 5 years and have a 4 yr child out of my wedlock. My married life had been full of problems from the first week of marriage - most of which can be summarized as dowry related harassment, physical and mental torture and cruelty. Now my husband and family have filed for divorce mostly on the grounds of cruelty and infidelity using false allegations to malign my character and false allegation to prove that I have been a bad daughter in law. …... I want to file a FIR and complaint in Women Cell regarding my jewellery and dowry related harassment. …. I want the child custody, monthly maintenance and share in husband's or in-law's property…"*

Our system analyzes the above text and discovers the domain dependent terms with the help of the CSN. For example, in the 2nd line, the system extracts the keyword '*marriage'* and associates the phrase '*full of problems*' with it. Also from the CSN, it can associate the following terms from the narration sentences. While doing so, it also consults the WORDNET synonym set to add a few more synonyms of those keywords depending on their parts of speech tag used in the search text. The system

creates a collection of sets containing *1…m* number of sets, each of which is again a set of *n* number of keywords. The cardinality of these sets appearing inside the superset will be ranging from *2….n* and the content of these sets will be constructed using the keywords in the search text and also taking the keywords from the conceptual network connections which are associated with them. In case, there are phrases in the search text appearing inside quotation marks, they are directly included in the collection set even if they are not found anywhere in the conceptual semantic network. A sample of such kind of collection of sets are presented below that can be generated from the above example search text S1.

[{*marriage, problem*}, {*marriage , dowry*}, {*marriage, 'physical torture'*}, {*Marriage, dowry, harassment* }, {*Separation, child custody}, {divorce, maintenance}*, {*Divorce, 'false allegations'*},{*marriage, endowment, harassment*}, {*Marriage, Dowry, mental torture*}, {*marriage, dowry, emotional abuse*}, {*marriage, dowry, 'verbal abuse', 'physical abuse', harassment*}, {*marriage, dowry, 'verbal abuse', 'physical abuse', harassment, divorce*}, {*marriage, 'physical abuse' , 'abusive marriage', cruelty, 'mental torture', separation*}]

The above collection of sets contains the possible candidate search queries that are ready to be passed to the background indexed data for information retrieval.

If the search text contains Named Entities (NE) which represents location or organizations, then those NEs are also added in the collection of sets. The system ignores the named entities relating to the names of persons because the names of persons appearing in the legal documents residing in the background may not have any significant importance compared to the narration of the legal aspects mentioned in the legal text. Because, background data was stemmed and stored in the index form, the possible set of search queries are also passed to the indexed data after stemming.

2.4 Ranking of the retrieved documents and selection of final set of documents

We have adopted the Lucene based searching techniques which uses a combination of Vector Space Model (VSM) [13] and Boolean Model (BM) [14]. In our case, documents are ranked and scored by VSM; only for those retrieved document which are approved by the BM. When passing the queries, each of the sets of queries inside the collection set of queries are sent separately and the returned set of documents are stored with their corresponding scores. VSM score of a document *d* for the query *q* is calculated using Cosine Similarity of the weighted query vectors *V(q)* and *V(d).*

$$Cosine\ Similarity\ (q,d) = \frac{V(q) \cdot V(d)}{|V(q)|\,|V(d)|}$$

Here, *V(q).V(d)* is the dot product of the weighted vectors, and *|V(q)| and |V(d)|* are their Euclidean norms.

Search queries having larger cardinality in terms of the containing keywords within them are given higher priorities when the ranking points are taken into consideration for the final set of documents. Because, the returned documents containing more significant keywords have the best chance of being more relevant than those containing lesser number of keywords. Finally, from the results set of each queries, we take the top 1000 highest ranked documents and build the final set of documents which are expected to be relevant to the search text.

3. EXPERIMENTAL VERIFICATION

For the experimental purpose, we have used the data set and queries from 'FIRE 2013 Ad-hoc retrieval from legal documents' track. The background data contained a set of small documents constituting over 3 Gigabytes in size. The documents contained verdicts from the Supreme Court, various acts of parliament. i.e. descriptions of situations in which legal assistance is required on two different domains namely 'Consumer Law' and 'Hindu Marriage & Divorce Law'. The number of documents containing such kind of verdicts were in excess of Hundred and sixty thousands for each of the domains.

There were 20 different search text comprising 10 from each of the domains for the document retrieval. All of the search texts for both of the domains were in the form a descriptive scenario. Our system used the proposed approach and tried to get the topmost 1000 document for each of the queries.

Our system managed to retrieve top 1000 documents for each of the queries from the respective domains. For example, the queries related to the consumer law legal documents were retrieved from the consumer law domain index. Similar operation was performed in the case of Hindu Marriage & Divorce Law queries. Our system also retrieved top 1000 documents for each of the queries from the combined domain index rather than the relevant domain.

4. CONCLUSION

Instead of analyzing either the background text or the search text and understanding the meaning of the content in depth, the presented approach in this paper attempts to discover a few significant keywords and there relation other concepts from the search text. Creating a domain specific Conceptual Semantic Network (CSN) and selection of significant keywords from the same is at the core of the system that we have proposed in this paper. We think, the creation of such domain specific network structure is feasible specifically under the legal domain as the number of terms or keyword phrases participating in the conceptual network is limited in number. Besides the selection of keyword phrases from the CSN, using the Wordnet synonyms of the keywords also helped the system to achieve wider retrieval coverage. The concept of forming a collection set of search queries has been found to be useful in this case.

Because of limited amount time that we had, we could not design the CSN as expressively as we wanted to and thus the retrieval results that we have achieved have lots of scope for improvement. This paper has been presented as a work of an ongoing research and we believe the proposed system can be further improved and presented as a functional system in domain specific Information Retrieval system with a narration based search text.

REFERENCES

[1] Singhal, Amit (2001). "Modern Information Retrieval: A Brief Overview". *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* 24 (4): 35–43.

[2] C.D. Manning, P. Raghavan, H. Schütze. Cambridge UP, 2008. Classical and web information retrieval systems: algorithms, mathematical foundations and practical issues.

[3] Anderson, J.D., & Perez-Caxrballo, J. (2001). The nature of indexing: How humans and machines analyze messages and texts for retrieval. Part 1&2. (Research and the nature of human indexing) Information processing and management 37 (2): 231-277.

[4] Frakes, William B. (1992). Information Retrieval Data Structures & Algorithms. Prentice-Hall, Inc. ISBN 0-13-463837-9.

[5] M.F. Porter, 1980, An algorithm for suffix stripping, *Program*, 14(3) pp 130−137.

[6] Knuth, D. E. (1997) [1973]. "6.5. Retrieval on Secondary Keys". *The Art of Computer Programming* (Third edition). Reading, Massachusetts: AddisonWesley. ISBN 0-201-896850

[7] Fire 2013, Forum for Information Retrieval Evaluation, http://www.Isical.ac.in/~clia/, Accessed on November 10, 2013

[8] OpenNLP Manual(http://opennlp.apache.org/ documentation/1.5.3/manual/opennlp.html , Accessed on November 10, 2013.

[9] Kristina Toutanova and Christopher D. Manning. 2000. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, pp. 63-70.

[10] Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of HLT-NAACL 2003*, pp. 252-259.

[11] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005),* pp. 363-370.

[12] George A. Miller. 1995. *WordNet: A Lexical Database for English,* Communications of the ACM Vol. 38, No. 11: 39-41.

[13] G. Salton, A. Wong, and C. S. Yang (1975), "A Vector Space Model for Automatic Indexing," *Communications of the ACM*, vol. 18, nr. 11, pages 613–620. *(Article in which a vector space model was presented)*

[14] Lashkari, A.H.; Mahdavi, F.; Ghomi, V. (2009), *A Boolean Model in Information Retrieval for Search Engines*, doi:10.1109/ICIME.2009.101