From Words to Concepts in Text Mining

JDIM
Journal of Digital
Information Management

Saleh Al- Zahrani
Information Systems Department
Faculty of Computer & Information Science
Imam Mohammad Bin Saud University
Riyath
Saudi Arabia
Dr_Saleh@hotmail.com, Sgzahrani@imamu.edu.sa

ABSTRACT: *In the text mining process the interface between the natural language text and content reflecting terms is essential for the successful indexing process. The traditional interfaces such as thesauri and lexicons have limitations for web content processing. The recent lexical nets have promises, but their efficiency needs to be tested. The current work measures the use of word relations specified by lexical nets in a large test bed. The results correlate with similar recent past studies as not all specified word relations in the lexical nets are semantically rich in expressing the conceptual relations between given words. The results call for applying more heuristic approaches for text mining.*

Categories and Subject Descriptors
H.3.1 [Content Analysis and Indexing]**:** Indexing methods **H.3.3** [Information Search and Retrieval]**:** Retrieval Models

General Terms
Text Mining, Web Content Processing, Information Retrieval

Keywords: Text Mining, Text processing, Word Relation, Concept indexing

Received 19 Dec. 2005; Revised and accepted 22 Mar. 2006

1. Introduction

Text mining processes mostly employ the keyword searches where keywords are extracted from text or using standard thesauri. Key word employment depends on either detecting the significant content reflective words or high frequent words. However, the key word approaches do not ensure semantic retrieval as the confinement of concepts to words is away from reality. Keyword enhancement is applied using term weighting and relevancy ranking in some web processing and search systems. The strength and short comings of these approaches are reviewed in [1].

2. Related Work

Text processing using key words was the focus of information retrieval research since 1950s. Traditionally, the indexing process (2) relied on the use of controlled vocabularies in order to achieve greater consistency and to improve the indexing quality (3).

Text-based search tools generally rely on some form of string matching, which may find difficulty with respect to misspelled words [4] and morpheme or cross-lingual related problems [5]. Besides, many other issues are raised and discussed in the recent web content processing.

The crux issue in text processing is to find the content reflective words. Tags are employed widely to extract the key words from text and a score of mechanisms found to be useful. The Parsers are used to tag phrases while marking the key terms in some experiments such as in MURAX.[6] Texts are fragmented to manageable units in the fact retrieval system of Cooper who used the text fragments in a small document collection which confirmed or denied a query statement [7.] Many systems currently employ parsers to locate and type the important phrases within the larger texts.

Instead of tagging words, sentences are employed in the CIQUEST [8] where the sentences were ranked using the criteria such as - presence of a key phrase in the sentence, a high number of common terms, and the position of the sentence as found in the document.

3. Nature of text mining

Text mining has problems despite the fact that the text is relatively permissible to recognize and manipulate text strings. It is somewhat difficult to master the gamut of natural language terminologies, and hence apparently, indexing proves to be a formidable challenge.

It is hard to find thesauri that cut across traditional disciplinary lines, as they limit the deployment of terms for processing. The term matching activity is prone to many misinterpretations. Classical Information Retrieval systems have addressed the problem using thesaurus where words are placed at varying levels. However, in the later stage it was found that the use of these operators alone are not sufficient due to the fact that there are words which have equal or near equal relation and linguistic attributes of words are not confined to selected relations revealed by the controlled vocabularies. [9] In the recent past, the lexical tools such as WordNet(10) and Lexical Free Net (11) were introduced and which found to address many information retrieval problems relating to terms processing. They enable to layer the unstructured information otherwise the content remains less useful. (12)

A concept is represented in the text using many terms and the usage is restricted in the text to a single term even the options are more. This leads to the non-retrieval of warranted term due to the differences in the use of words for querying and text representation. This problem is recognized as the most predominant problem and the classical information retrieval techniques addressed in different ways.

Key words are either creator-driven or crawler-identified which is based on content bearing words from large texts. Most of them rely on term frequency where terms are given weight using many mechanisms. This way could become an ideal system, if concepts are represented using unique band of words. In the natural language, different terms represent concepts which need to be unified using lexicons.

A major unresolved issue in text mining is that the intermediary tools such as word databases, glossaries, thesauri and others provide a kind of match between the query terms and terms for indexing, and do not perform well in identifying the context or word relations. Among these tools, the word and lexical nets have introduced enhanced features and offer promises for text indexing. Through this exercise, we estimate the efficiency of the kind of word relations identified by Lexical Free Net, a lexicon tool that identifies possible word relations. The lexical free net scores more than thesauri and glossaries as i) it is interdisciplinary and multidisciplinary in coverage; and, ii) it specifies a possible kind of relations for a given set of words.

However, during our initial applications, we found a few problems while extracting word relations. The base level problem is that not all relations are applicable for a given word in the text and application of all relations while indexing will lead to distorting results. Hence we would like to test the degree of effectiveness of using lexical free net in indexing process.

For the current experiment we built an interface by extending the one of Jacobs [9]. The interface we built has the following elements:

1. Texts and queries have strings where the strings have both single words and associative words.
2. The text and query words are recorded in the interface database.
3. The interface process query words and bring searchable words which enter into search process

without second level querying by users.

4. Lexical net offers the words in terms of paths for query words with word relations.

Based on the above premises, we have used the following algorithm.

```
QuerySet = LFNet files
QuerySet = Feed (QueryTerms) /* Listing paths */
QuerySet = Pre-process(QuerySet) /* Attribute selecton,
           synonym conversion */
QuerySet = sort(QuerySet) /* path analysis for query terms */
For each word do
Cluster all one-Word queries
Repeat
Cluster attributes to query /* Use LFNet to place */
Use to choose two or more words with one keyword being
        different
Produce a list
Insert the list into QuerySet; remove all irrelevant from QuerySet
Until no irrelevant can be found
Output all semantic value attributes
Count/*
```

Figure 1. Path mining algorithm

4. Relations identified in Lexical FreeNet

The modern lexical nets have enhanced specifications of word relations. The Lexical FreeNet allows 16 relations among the words. These sixteen attributes bring returns score of paths during word-relation identification. The searchable words return the many paths where the paths are analysed using the algorithm described.

5. Search

For the current work, we have keyed 128 selected individual words which have returned a large number of paths. Each path is checked for relevance with query word. The table 1 shows the view on the query terms and path information.

The basic part of the current first phase records the path returned with the linguistic approach of the LexicalNet thesaurus. More specifically, we investigated the relationship between feed words and term specificity for identifying concept relevance. The question to be addressed was, given a pair of term/concepts that have been found to be related, how does one determine which is more relevant? The aim was to measure on a large scale the accuracy of retrieved term sets for given word pairs taken from LexicalNet.

Semantic measurements are examined by matching the retrieved paths with established thesauri. Path relevance for a dataset was then estimated by averaging the number of terms retrieved for each term from established thesauri and glossaries. In the experimental design, it was assumed that the commonest sense of a term accounts for the great majority of the returned paths and that this would correspond to the retrieval of commonest sense in LexicalNet. Among the semantic relations identified using glossaries and thesaurus, heuristic experiments were initiated.

The single term querying in the lexicon has brought 92451 paths where approximately one third was found to be related with the query context and multiple query terms have produced one fourth of the total paths returned which have semantically relevant paths. Our results are similar to Jacobs et al [9] who obtained the results corresponding to our experiment when they did a small scale testing. This lead to the caution that not all paths are semantically rich and the application of all paths for indexing will not produce results with precision.

6. Index Creation Paths

Not all the returned paths would be related to the text files as the lexicon identify all possible relations a word likely to posses. The analysis of semantic richness is carried out now.

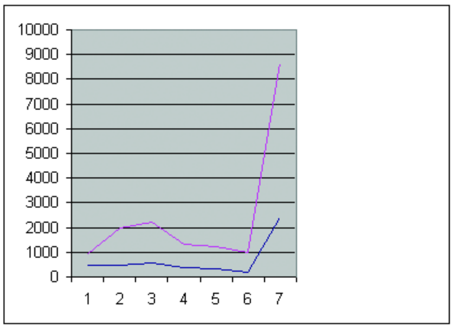

Figure 2. Total Paths versus semantically rich paths

In the figure 2, we have presented the total paths returned in general for terms and the semantic rich paths for selected relations. The lexical net has given many possible relations; not all have equal semantic relevance. We have selected six relations for testing such as Synonyms, Trigger, Generalises, Specialise, Comprise and Part. Even with progress on semantic measurements, still we lack concrete and acceptable method for semantic detection as relevance is user-perceptive and independent of words. Thus, we employed manual semantic detection of returned paths and hence the analysed path population is insignificant. However, we have given enough evidences as all relations specified in Nets are not equal. Glossaries or Nets are independent of contextual specifications.

The information retrieval experiments in the past documented that all related terms of a given word do not match with users' expectation. However, the context of the words employed in texts may differ. The lexical freenet used as many as sixteen possible word relations. It is true that not all attributes are applicable for the text indexing. And also, a possible improvement of the lexical freenets is that the word relations are not confined to those lexical nets identified ones.

We have measured the retrieval effectiveness of the results by measuring the recall and precision of the data in the table 3. Expect synonyms relations, the results for other relationsare not encouraging.

Query Terms	Terms Used	Returned Paths	Mean Path attributes Available	Semantic useful paths	Percentage of semantic useful paths
Single	928	92451	9.96	28654	30.99
Multiple	832	87333	10.49	21623	24.75
Pair	2342	25323	10.81	8959	35.37

Table 1. Query terms and paths returned

Measures	Synonyms	Trigger	Generalises	Specialises	Comprises	Part
Recall	0.86	0.68	0.59	0.28	0.42	0.41
Precision	0.72	0.31	0.24	0.57	0.45	0.19

Table 2 Relevance measures of path validity

The data on relations identified through the attributes to the query are significant in information retrieval design. The words for the six attributes are tested for the relevance measures - recall and precision. The scores are poor for five out of the six measures. Only for the synonyms attribute, it is high.

7. Summary and Conclusion

In this study, we have presented our prototype results with a large testbed. The ongoing experiments would produce concrete results and we do hope that the future text indexing exercises can use built interfaces. The use of any lexicon or controlled vocabulary or thesauri for processing activities without adhering to the semantic and context bearing terms may yield to poor precision. The use of semantic attributes enables to use the Index creating rather than browsable attributes.

If the entire path attributes are tagged with robust and consistent keywords, most searches let the users to unwarranted text. The work has proved that tagging by using semantic rich paths would annotate the text and ostensibly improve the performance of search tools that take tags into account. The study results offer promise for future research. The semantic rich paths when used to create tags, the tags become more scientific and functional. Meta tag standards for web pages were introduced way back in 1996, and other metadata standards for information retrieval date back to the 1970s. When, metadata is created by using well-defined semantic rich paths, the indexing process would certainly achieve perfection in information processing and retrieval.

Acknowledgement

The author expresses his thanks to Professor Daisy Jacobs and Professor Pichappan for the permission to use their basic testbed.

References

1. Jensen, B J., Spink, A., Bateman, J., Saracevic, T (1998). Searchers, the subjects they search and sufficiency: a study of a large sample of EXCITE searches. Proceedings of the WebNet'98, Orlando Florida.

2. Anderson, J.D., Pierez-Carballo, J (2001). The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part I: Research, and the nature of human indexing, *Information Processing and Management*, 37 (2) 231–254.

3. Svenonious, E (1986). Unanswered questions in the design of controlled vocabularies, *Journal of the American Society for Information Science,* 37 (5) 331–340.

4. Navarro, G (2001). A guided tour to approximate string matching, *ACM Computing Surveys*. 33 (1) 31–88.

5. Schulz, S., Hahn, U (2000). Morpheme-based, cross-lingual indexing for medical document retrieval, *International Journal of Medical Informatics*, 58, p.87-89.

6. Kupiec, J (1993). MURAX: A Robust Linguistic Approach For Question Answering Using An On-Line Encyclopedia, *In*: Proceedings of. the 16th annual international ACM SIGIR conference on Research and Development in Information Retrieval, 181-190.

7. Cooper, W.S (1964). Fact Retrieval and Deductive Question-Answering Information Retrieval Systems, *Journal of the AC*M, 11 (2) 117-137.

8. Joho, Hideo., Sanderson, Mark (2004). Retrieving Descriptive Phrases from Large Amounts of Free Text, *In*: Concept-based Interactive Query Expansion Support Tool (CIQUEST) edited by Micheline Beaulieu, Mark Sanderson, and Hideo Joho, http://ciquest.shef.ac.uk/Licrr149.pdf

9. Jacobs, Daisy., Srinivasaraghavan, S., Pichappan, P (2006). Text Mining Using Lexical Nets: An analysis of word relations, *In*: Fourth International Conference on Computer Science and Information Technology, April 5-7, 2006. Amman.

10.http:// www.wordnet.com

11. http://www.lfnet.com

12. Maina, Ernest Weke., Ohta, Manabu ., Katayama, Kaoru., Ishikawa, Hiroshi (2005). Semantic Image Retrieval based on Ontology and Relevance Model - A Preliminary Study, *Journal of Digital Information Management*, 3 (4) 227-230.

Dr. Saleh Al-Zahrani is assistant professor at Imam Mohammad Bin Saud University, Saudi Arabia. He is the head of Information Systems Department at the Faculty of Computer and Information Science. He has published several papers on information security and information retrieval in different conferences and journals. Moreover, he is active member in several computer science professional associations nationally and internationally.