Using Ontology to build News Network

M.S. Razali, A.Y. Yusufie, S.M.F.D. Syed Mustapha

Universiti Tun Abdul Razak, MALAYSIA
ahmadyusufie@gmail.com

ABSTRACT

One of the main activities in the knowledge sharing is to search and retrieve textual document. Traditional searching methods use user-specified keywords to search for documents. The common problem with this method is that the retrieved documents are not the ones that they are actually looking for even the searching is based on user-defined keywords The proposal in the research work is to build a well-defined domain where semantic relationship can be defined among the text documents in the repository to enhance the searching and retrieval performance. Reuters news is chosen as the domain where the ontology that defined the relationship is established to address the synonymy and polysemy problems. The ontology uses keywords to quantify the relationship strengths and labels the qualitative semantics. The ontology structure is a network of documents that is arranged based on hierarchy. This paper discusses the implementation of the document ontology which is applied to Reuters news corpus where the retrieval performance is measured based on the recall and precision.

Keywords
Semantic, Synonymy, Polysemy, Ontology

1.0 INTRODUCTION

The cyber news user community nowadays needs a searching system which can give them what they really want, rather than being told of what they should take. Hence, this development of news network using ontology.

Historically, ontologies arise out of the branch of philosophy known as metaphysics, which deals with the nature of reality.This fundamental branch is concerned with analyzing various types or modes of existence, often with special attention to the relations between details and generalities. The traditional goal of ontological inquiry in particular is to divide and to discover fundamental categories.

During the second half of the 20th century, philosophers extensively debated the possible methods or approaches to building ontologies, without actually building any elaborate ontologies themselves. By contrast, computer scientists were building some large and robust ontologies.

Since the mid-1970s, researchers in the field of artificial intelligence have recognized that capturing knowledge is the key to building large and powerful AI systems. AI researchers argued that they could create new ontologies as computational models that enable certain kinds of automated reasoning. In the 1980s, the AI community began to use the term ontology to refer to both a theory of a modeled world and a component of knowledge systems themselves (Liu, Ozsu & Springer-Verlag, 2008).

In the early years of the 21st century, the interdisciplinary project of cognitive science has been bringing the two circles of scholars closer to. But there are still many scholars in both fields uninvolved in this trend of cognitive science, and continue to work independently of one another.

In computer science and information science, an ontology is a formal representation of a set of concepts within a domain and the relationships between those concepts themselves. It is used to reason about the properties of that domain, and could be used to define the domain.

The purpose for this research derived from the realization of the lack of good searching method for textual information for news retrieval. A generic parametric algorithm is developed that analyzes the text document and determines the relevancy based on certain controlling values. Controlling values are values that can change overall output of the system. It is a threshold value where the algorithm will terminate once the threshold value is reached. For example when we are picking up keywords from an article, the words are only considered as keywords if the frequency is at or above a certain value which we call a threshold. This value is one of the controlling values we used in the system. The precision and recall of information retrieval can be better balanced by adjusting the controlling values.

In this paper, each news articles from Reuters corpus are incorporated into an ontology and structured as hierarchies of different layers using a tool. The development of the ontology from the group of keywords will be introduced. Semantics and strengths will be added to make the searching of documents based on the ontology to be more efficient.

2.0 RELATED WORKS

One of the main problems in text retrieval is how to deal with multiple terms that refer to the same concept, which is what we call synonymy (Edmonds & Hirst, 2002; Jones, 1986]. The developers of text retrieval systems have tended to solve it with query expansion using controlled vocabularies containing synonym lists or classification hierarchies. The retrieval accuracy may improve but the technique requires manual effort which can be costly.

A paper (Bradshaw, Scheinkman & Hammond, 2000) however have pointed out that users do not always submit discrete queries to information retrieval systems, even though detailed queries are often necessary to get highly-specific search results.

They (Bradshaw et al., 2000) then have come up with a way to index documents so that a query will give high-quality search results even though the query is not very discrete. It is that research documents are indexed according to how they were referenced in other documents. This reference-based query expansion method was created based on the observation that in research papers, the text surrounding a reference is usually a description of the information the referenced documents provides. By doing this, the indexing is more powerful because references pair has concise descriptions with the documents that contain that information.

However, one of the problems that the system has is a lot of recent articles are not found. This is because for the articles to be indexed, it must be referenced by another article first. An article may have to wait for years for it to be referenced by another article, therefore to be indexed by this system.

The second main problem in text retrieval is what we call polysemy, where similar terms can belong to different concepts. Developers of the early information retrieval systems chose to ignore this problem. Relevance is simply based on word similarity in those days (Bart, Robertson & Sorace, 1996).

By using this method, a given system can retrieve many relevant documents, but at the same time also taking irrelevant ones. The words that are the same but have different meanings are also taken. The worse part is that documents which are relevant but does not have the same words will be omitted.

The result of ignoring this polysemy problem is low precision and low recall. Low precision will cause a lot of new problems in latter phases, while low recall is what we were trying to avoid in query expansion method (Burnard, 1995).

Jing and Tzoukerman of Columbia University and Bell Labs (Jing & Tzoukerman, 1999) made an assumption that a given word has a dominant meaning in a given document. They then made up context vectors which consist of the frequencies of the words that occur within a window of 10 words surrounding the target word. The more frequent a given word appears within the window of the target word, the more important it is in the vector.

Now that the context vectors consisted of single words, target words will be paired from one vector to another to see how close and semantically related they are. They will be compared using co-occurrence frequency. If the terms in one context vector have strong co-occurrence relationships with the terms in another context vector, then the target words are likely to be semantically related (Corley, Corley, Keller, Crocker & Trewin, 2000; Cowart, 1997)

The calculation of word pair co-occurrence strength makes it possible to calculate the similarity or difference between context vectors. This context vector theory is an important contribution to a great problem in information retrieval. It gives us the way of how to find the optimal balance between precision and recall.

However, context vector method proposed here is not generic. It can only work for a given corpus. This is why controlling values that are adjustable are allowed in the proposed algorithm.

3.0 ONTOLOGY FOR NEWS NETWORK (MODIFIED CONTEXT VECTOR)

From a set of news corpus, we will be building an ontology automatically using a tool. Therefore, context vectors must be established first. Only then the documents will be associated with the vectors which they belong to. Only then, hierarchies could be built, depending on the need of it. Different hierarchies will be built for different vectors.

Context vectors here are groups of keywords. Each keyword group consists of keywords which are found in a few documents, depending on how many documents that are set initially to the system to establish keyword groups.

3.1 Keywords Extraction

Every word that has been stemmed from each article together with its frequency will be stored into a database. The frequency will then be normalized. Each words which are higher than the threshold (varies according to news provider), are considered as the keywords for that given article. We have chosen "0.8" as the threshold for the experiment data Reuters news corpus, because after some initial experiments we found out that the number is the most suitable. If we use any smaller number of threshold, we will retrieve a lot of irrelevant keywords.

3.2 Keyword Group Establishment

Keywords from each article are compared from keywords from different articles in the set. If there are few keywords which appear in a few documents, then that few keywords will be put in a group called the keyword group. Each document that has the same keyword which belongs to a group is associated to that keyword group. We have used "2" as the number of documents a keyword should be in for it to become a keyword group. The number is enough for the keywords to be strong according to the experiment data, because a high threshold was set.

3.3 Hierarchy

Keyword groups which have too many documents associated to them are broken into a hierarchy (layers). The layers of hierarchy will be based on how many documents attached to a group. If there are too many documents, then the next layer has to be established. Using this experiment data, "3" was used as the number of documents a hierarchy should have in 3 level. This is due to many documents were retrieved in the experiment.

3.4 Strengths and Semantics

Strengths and semantics of the documents will be added to every document. This will be done by comparing the associated documents with a database of strengths and semantics. The strengths will be based on the weight of the each document, which is the average weight of all keywords from each document. The weight of a keyword is determined by the normalized frequency. The semantics will be based on the keywords. This is a very crucial process as it will effect the searching of the news articles later.

4.0 PREPARATION OF DATA

Input data were downloaded from Reuters database to the system's database. It is pre-organized before the system is run. Once the system is run, the data will become the output according to what the system is asking and the structure of how it was stored after the pre-organization.

5.0 EVALUATION

A Reuters news corpus consists of 925 articles was used. Each keyword will be put into groups. So there will be 925 groups initially. Then the keywords in those groups will be compared.

From Figure 1, we show an example of three similar keywords being shared between document 1 and document 2.

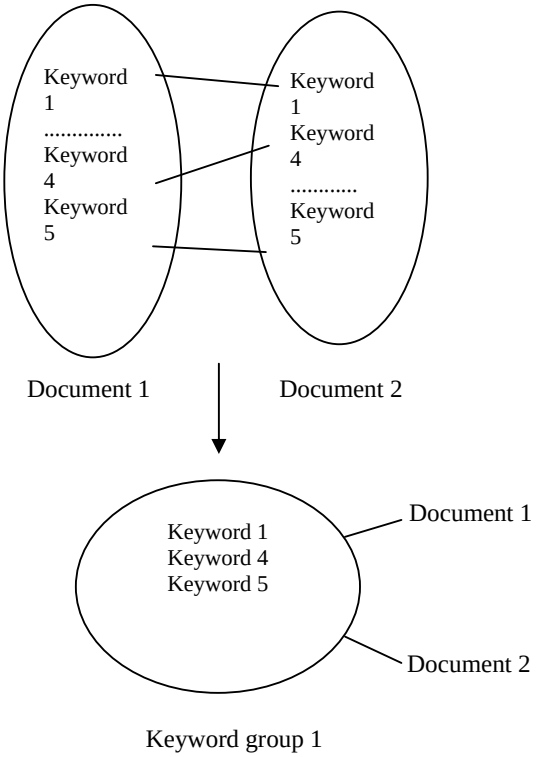

Figure 1: The establishment of keyword groups

Figure 2 shows the flowchart of how a keyword group is established. The controlling values "0.8", "2", and "3" can always be changed in the tool that has been created by us. This is to keep the generality of the best result depending on which news corpus being used.

6.0 RESULTS AND DISCUSSION

The final result was 1097 keywords, which is in average 1 keyword per article. 134 keyword groups were established, which is in average 7 documents per are attached to a keyword group. Hierarchies were built.

7.0 CONCLUSIONS AND FUTURE WORK

We have developed a new method for searching which is more efficient and user-friendly. However, we still would like to explore other different related methods and expand our theories.

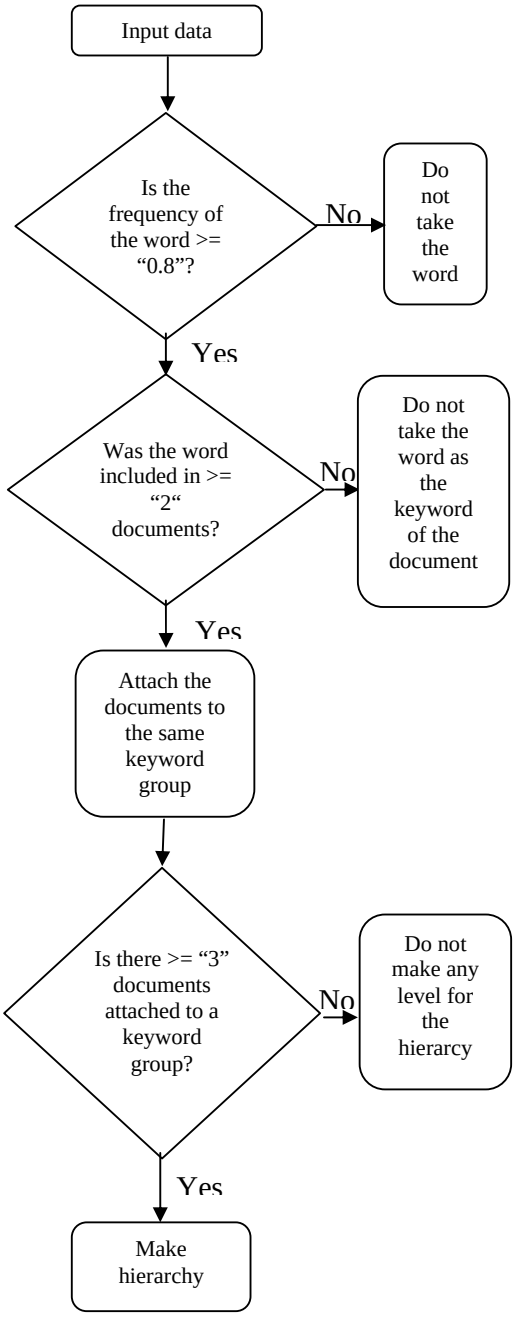

Figure 2: Flowchart of how a keyword group is established

REFERENCES

Bard, Robertson, & Sorace (1986). Magnitude estimation of linguistic acceptability. Language. *72*(1), 32–68.

Bradshaw, Scheinkman, & Hammond (2000). Guiding People to Information: Providing an Interface to a Digital Library Using Reference as a Basis for Indexing. *Proceedings of the 5th International Conference on Intelligent User Interfaces*, pp.37-43.

Burnard (1995) Users Guide for the British National Corpus. *British National Corpus Consortium, Oxford University Computing Service*.

Corley, Corley, Keller, Crocker & Trewin (2000). Finding syntactic structure in unparsed corpora: The Gsearch corpus query system. *Computers and the Humanities.*

Cowart (1997). Experimental Syntax: Applying Objective Methods to Sentence Judgments. *Sage Publications, Thousand Oaks, California*.

Edmonds & Hirst (2002). Near- Synonymy and Lexical Choice. *Computational Linguistics, 28*(2), pp.105-144.

Jing & Tzoukerman (1999). Information Retrieval Based on Context Distance and Morphology. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 90 – 96.

Jones (1986). Synonymy and semantic classification. *Edinburgh University Press, Edinburgh, Scotland*.

Liu, Ozsu & Springer-Verlag (2008). *Encyclopedia of Database Systems.*