Keyword Extraction from Meeting Documents for Search and Retrieval

Caslon Chua, Clinton Woodward
Swinburne University of Technology
Faculty of ICT, H39 PO Box 218, Hawthorn, Victoria 3122, Australia
{cchua, cwoodward}@swin.edu.au

ABSTRACT

Information plays an important role for management to make sound decision for the organization. Organization today keeps documents in electronic format; however it appears that aside from the possible, well-organized folders in a disk and well-meaning filenames given to documents, searching documents based on contents is difficult. In addition, most if not all document search requires the user to recall exact phrases in the document for search and retrieval. In an organization where administrators changes over time, decisions, recommendations, suggestions and other important things that were recorded formally in a document may remain unknown to new administrators. New administrators may have to exert pro-active efforts to retrieve previous decisions before making a new one, due to possible contradiction or repetition. As the volume of document increases, search and retrieval become tedious and difficult. This study presented an approach to parse and analyze meeting documents to extract keywords in preparation for indexing and clustering.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Content Analysis and Indexing – dictionaries, *indexing methods, linguistic processing*

General Terms

Algorithms, Languages

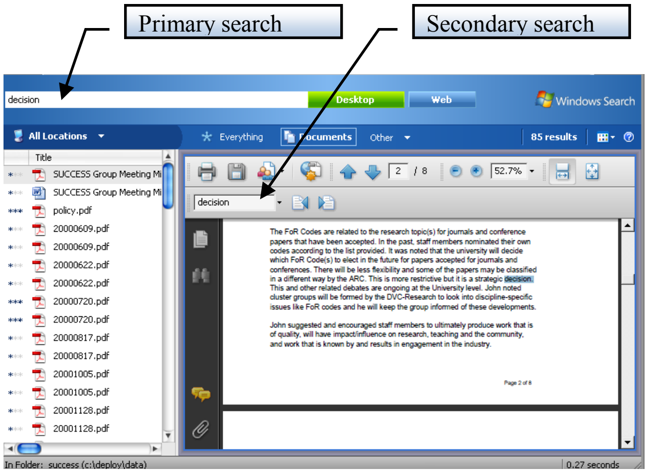
Keywords

information search, information retrieval, document clustering.

1. INTRODUCTION

In an organization, conducting meetings is a way in which staff, managers, and administrators come together and discuss important issues concerning the organization with the purpose of making decisions, recommendations and suggestions among others. These meeting discussions are recorded as minutes of the meeting which served as official documentation for staff, managers and administrators to refer to. Depending on the organization, a number of meetings may be conducted over a period of time. This means the number of meeting documents grows over time. Thus the ability to perform search and retrieval is important in order to bring relevant and important information for reference.

While most organizations keep meeting documents in electronic format, searching through them can be quite tedious. The conscious effort of storing these documents in chronologically named directories and filename does not provide one with an idea of what was discussed and recorded in the meeting document. While there exists a number of desktop search applications, these search applications treat each document as a single entity. The search would lead the user to a document containing the keywords being searched for; however, users will then have to refine their search within the document, and probably requiring the user to browse through the document before realizing the usefulness of the retrieved document. For instance, Windows Desktop Search in Figure 1 shows the primary search result of “decision”. It directs the user to the meeting document that contains the word “decision” after the desktop search, and the user will have to refine the search within the preview using the secondary search.



The FoR Codes are related to the research topic(s) for journals and conference papers that have been accepted. In the past, staff members nominated their own codes according to the list provided. It was noted that the university will decide which FoR Code(s) to elect in the future for papers accepted for journals and conferences. There will be less flexibility and some of the papers may be classified in a different way by the ARC. This is more restrictive but it is a strategic decision. This and other related debates are ongoing at the University level. John noted cluster groups will be formed by the DVC-Research to look into discipline-specific issues like FoR codes and he will keep the group informed of these developments.

John suggested and encouraged staff members to ultimately produce work that is of quality, will have impact/influence on research, teaching and the community, and work that is known by and results in engagement in the industry.

Figure 1. Document search using Windows Desktop Search

This type of search requires the user to know the exact keyword that must be used in the search. Furthermore, keywords with different inflection will be missed in the search. For example, words such as decisions, decide and decided will not be located. In addition, if the text in the document uses synonyms to imply a decision is made such as resolution, settles on, and chosen, that document will also not be located.

8th National Natural Language Processing Research Symposium, November 24-25, 2011, Manila, Philippines.

This study is on the development of an approach to parse and analyze minutes of the meeting documents in order to extract keywords that facilitate an electronic search and retrieval given a user query. The goal of the study is to compare different approaches in keyword extraction from the meeting document that can represent key thoughts of the topic which are then used in a search engine tool for indexing and clustering.

The study defines key thought as the action item being emphasized in the document, such as decision, recommendation, or suggestion.

2. Related Studies

Keywords may serve as a dense summary for a document that lead to improved information retrieval, or an entry to a document collection [6]. In creating meeting documents, keywords are rarely, if at all, assigned to the document, thus developing an automated system to generate keywords for the document would be helpful. Automatic keyphrase extraction is defined as the automatic selection of important topical phrases from within the body of a document [12]. When keywords are used in a search engine, users are able to make the search more precise. A document search that matches a given keyword will result to a smaller, higher quality list of hits than a search for the same term in the full text of the documents [12].

Many studies on keyword extraction are aimed at facilitating information retrieval. There are four categories on keyword extraction methods. These are simple statistics, linguistics, machine learning and hybrid approaches [14].

2.1 Simple Statistics Approaches

These methods are simple and independent of the language and domain of the document. It uses statistical information on the words in the document to identify the keywords. Cohen proposed an approach to extract highlights based on representing the text by its n-gram counts and the document is represented by a vector detailing the number of times each sequence was observed [1]. Luhn proposed a statistical method towards supporting automatic encoding of documents for future information retrieval [7]. Salton et al. proposed the use of discrimination value analysis to rank the text words in accordance with how well these words discriminate the documents in the collection from each other [9]. This is referred to as term frequency–inverse document frequency weight. Other term weighing approaches were also proposed by Salton [10]

2.2 Linguistics Approaches

These approaches are based on the linguistics feature of the word, sentence and document. Hulth experimented on term selection approaches that include noun phrase (NP) chunks and terms matching any set of part-of-speech (POS) tag sequences [6]. Ercan et al. describes a keyword extraction method that investigates the benefits of using lexical chain features [2]. Plas et al. worked on automatic keyword extraction from spoken text using lexical resources [8].

2.3 Machine Learning Approaches

These approaches employed supervised learning from examples to extract keywords. The machine learning starts with a set of training document to learn a model. The gained knowledge from the model is then applied to find keywords from new documents.

In machine learning approaches, researchers explored key phrase extraction aside from keyword extraction. Turney described a hybrid genetic algorithm for keyphrase extraction [12]. Frank et al. described a procedure for keyphrase extraction based on the naive Bayes learning scheme [3].

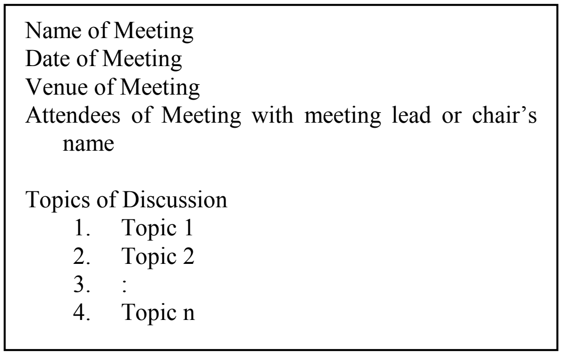
2.4 Hybrid Approaches

These approaches involve the combination of the above three approaches in extracting keywords. It may also employ some heuristic knowledge on the document in which the keywords are to be extracted. These include structure of the document, domain of the content, and language. Huang et al. described an algorithm that treats each document as a semantic network which holds both syntactic and statistical information [5].

3. System

3.1 Document Analysis

This study assumed that a partial structure exists in the minutes of the meeting documents as illustrated in Figure 2. Items or topics discussed in a meeting are written as paragraphs that are numbered with the possibility that the mentioned item or topic is again discussed or resolved in succeeding documents.



Name of Meeting
Date of Meeting
Venue of Meeting
Attendees of Meeting with meeting lead or chair's name

Topics of Discussion

1. Topic 1
2. Topic 2
3. :
4. Topic n

Figure 2. Meeting document structure

Given this structure definition, items or topics are treated as sub-documents by the study. Each sub-document is analyzed to allow extraction of keywords that can be used in clustering of the sub-document across the document collection.

3.2 Keyword Extraction

In the document collection used in the study, no keywords were provided by the author. The study tested keyword extraction using both the statistical and linguistic approach. Common stop words were excluded in both approaches. Extracted keywords from both approaches were mapped independently to the collection of sub-documents.

Figure 3 illustrates the mapping created by the extracted keywords to the sub-documents which is used to cluster the documents. It is assumed that there will also be cases that more than one keyword are extracted from a sub-document.

In implementing the statistical approach, exact text are extracted for use as keywords. Selection is based on the frequency count of the text. For the linguistic approach, keywords are further classified as key thoughts or subject depending if the extracted words are verbs or nouns.

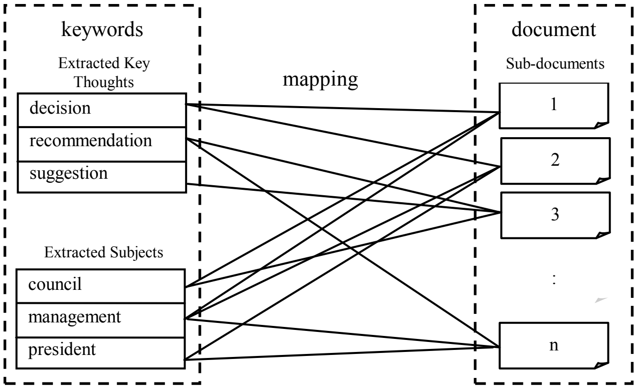


Figure 3. Extracted keywords and sub-document mapping

3.2.1 *Statistical Approach*

In the statistical approach, frequency count was applied on the document. Four types of frequency counts were performed on the document collection. These are frequency count within each sub-document with and without word stemming, and frequency count within the document with and without word stemming.

In frequency count within the sub-document, words that had a frequency count of more than one within the sub-document are classified as keywords. In the event that all words have a frequency count of one in the sub-document, the first word is automatically extracted as the keyword.

In frequency count within the document, words in the sub-document that had a frequency count of more than one in the document are classified as keywords for the sub-document. In the event that all words in the sub-document only occurred once throughout the document, the first word is automatically extracted as the keyword.

3.2.2 *Linguistic Approach*

In this approach, each sub-document is processed by a part-of-speech (POS) tagger developed by Stanford Natural Language Processing Group [11]. Words tagged as nouns and verbs are then extracted to represent keywords. Consecutive sequences of nouns or verbs form key phrases.

For example, given the sentence

It was noted that the university will decide which FoR Code(s) to elect in the future for papers accepted for journals and conferences

The following part-of-speech tag is generated.

It/PRP was/VBD noted/VBN that/IN the/DT university/NN will/MD decide/VB which/WDT FoR/NNP Code/NNP -LRB-/-LRB- s/PRP -RRB-/-RRB- to/TO elect/VB in/IN the/DT future/NN for/IN papers/NNS accepted/VBN for/IN journals/NNS and/CC conferences/NNS ./.

With common stop words excluded, Table 1 lists the extracted keywords based on the above sentence which are classified under noun and verb.

"FoR Code" is extracted as a key phrase. Moreover when extracting nouns, various tolerance levels were used that allowed prepositions and/or articles between two nouns. For example, if the text contains "Office of the President", it is also extracted as a single key phrase.

Table 1. Part-of-speech keyword extraction

Nouns	Verbs
university	noted
FoR Code	decide
future	elect
papers	accepted
journals	
conferences	

3.3 Keyword Association

In addressing the problem of searching documents using non exact keywords, keywords are further expanded by mapping it to its synonyms. However, keyphrases are not included in the expansion process. The study used the synonyms list from WordNet [13]. Figure 4 shows that additional keywords can also be used to search for the term "decisions".

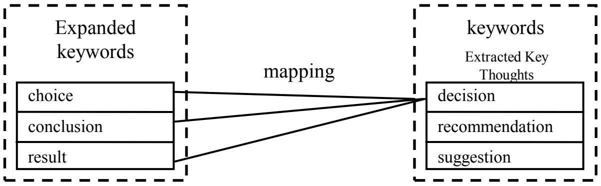


Figure 4. Expanded keywords list

4. System Test Results

4.1 Document Analysis

In testing the system, 103 documents were used in the collection. A rule-based system was implemented to analyze the each document. 98% of the documents were successfully parsed into sub-documents. Excluding the meeting information, each document contains an average 22.23 sub-documents and 289.38 words. In terms of limitations, the system was not able to recognize charts and tables found in 1 document. In addition, bullet points used within a sub-document is interpreted as a paragraph.

4.2 Keyword Extraction

In testing the system, keywords extractions were performed on the 103 documents collection. Table 2 summarizes the average number of keywords extracted from each sub-document.

Table 2. Keyword extraction count

Keyword Extraction Approach	Keywords per sub-document
Statistical Approach	
Frequency count per sub-document without stemming	1.28
Frequency count per sub-document with stemming	1.73
Frequency count per document without stemming	3.1
Frequency count per document with stemming	3.5
Linguistic Approach	
Part-of-speech Verb Tag	3.24
Part-of-speech Noun Tag	9.12*

* includes key phrase

In the statistical approach, the per sub-document statistics result to 94% of the candidate keywords have a frequency count of one.

Given that the first word in each sub-document will be selected by default, the keyword is not representative of what the sub-document is discussing. Furthermore, while there is an improvement in the per document statistics, the selected keyword represents the document more than it represents the sub-document. In the linguistic approach, the part-of-speech verb tags was able to capture the action described in the sub-document, and the part-of-speech noun tags represents the people or entity involved. Thus, in terms of quality of keyword extraction, using the linguistic approach resulted to better representation compared to statistical approach.

4.3 Keyword Association

Given that the documents do not have keywords assigned to it, 13 participants were invited to evaluate the quality of the associated keyword mapped to the sub-document. These participants have at least a bachelor's degree and are aged from 20 to 45. 85% are well versed in the use of search engines. Figure 5 shows that the quality of association is rated as good or very good by at least 36% of the participants compared to 29% who rated the association as poor or very poor.

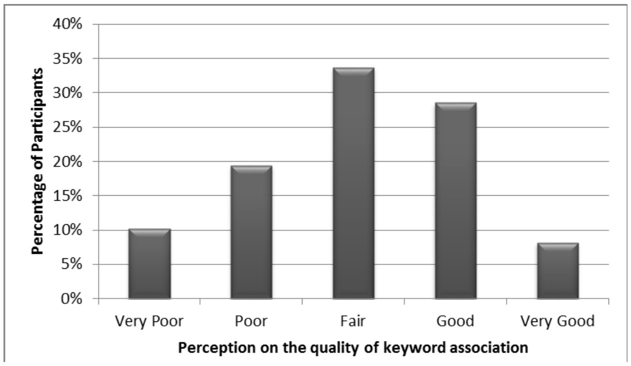


Figure 5. Quality of keyword association

In the evaluation, not all synonyms association were rated as good or very good, unlike association using actual key phrases from the document.

5. CONCLUSION

In comparing different extraction approaches, results showed that linguistic approach is able to extract keywords and key phrases that represent key thoughts of a sub-document in an unsupervised set up. Other features of the extracted keywords will be explored in future work as described in a study on text summarization extractive techniques [4]. This may include the incorporation of weight assignments to the extracted keywords. In addition, machine learning approach will be considered in the future.

Given that the manual assignment of high quality keywords is expensive, time-consuming, and error prone [14], the significance of systems that automate this process is important. The contribution of this study showed that aside from extracting keywords using linguistic approach, synonym association can also be used. However additional evaluation of synonym association will have to be conducted to explore techniques in synonym selection that can improve the quality of association.

In terms of application, these extracted and associated keywords can be used to enhance indexing and clustering used in search and retrieval. Furthermore, to explore ways of clustering, subsequent work will consider the use of hypernyms and hyponyms in keyword association. This will enable the creation of generalized or specialized clusters of sub-documents.

6. ACKNOWLEDGEMENTS

The authors would like to thank the participants of our experiment for their time and feedback. Thanks also go to the anonymous reviewers for their helpful comments.

7. REFERENCES

[1] Cohen, J.D. 1995. Highlights: Language and Domain-independent Automatic Indexing Terms for Abstracting. In *Journal of the American Society for Information Science*, 46, 3, 162-174.

[2] Ercan, G., Cicekli, I. 2007. Using Lexical Chains for Keyword Extraction. In *Information Processing and Management*, 43, 6, 1705-1714.

[3] Frank, E., Paynter, G. W. and Witten, I. H. 1999. Domain-Specific Keyphrase Extraction. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, (Stockholm, Sweden, Morgan Kaufmann), 668-673.

[4] Gupta, V. 2010. A Survey of Text Summarization Extractive Techniques. In Journal of Emerging Technologies In Web Intelligence, 2, 3, 258-268.

[5] Huang, C. Tian, Y., Zhou, Z., Ling, C. and Huang, T. 2006. Keyphrase Extraction using Semantic Networks Structure Analysis. In *Sixth International Conference on Data Mining,* (Hong Kong, Dec. 18-22, 2006) ICDM '06, 275-284.

[6] Hulth, A. 2003. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 Conference on Emprical Methods in Natural Language Processing*, (Sapporo, Japan, 2003).

[7] Luhn, H. P. 1957. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. In *IBM Journal of Research and Development*, 1, 4, 309-317.

[8] Plas, L., Pallotta, V., Rajman, M. and Ghorbel, H. 2004. Automatic keyword extraction from spoken text. A comparison of two lexical resources: the EDR and WordNet. In *Proceedings of the 4th International Language Resources and Evaluation*, European Language Resource Association.

[9] Salton, G., Yang, C. S. and Yu, C. T. 1975. A Theory of Term Importance in Automatic Text Analysis, In *Journal of the American society for Information Science*, 26, 1, 33-44.

[10] Salton, G. and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval, In *Information Processing & Management*, 24, 5, 513 – 523.

[11] Stanford Natural Language Processing Group, Stanford Log-linear Part-Of-Speech Tagger, http://nlp.stanford.edu/software/tagger.shtml

[12] Turney, P. 1999, Learning Algorithms for Keyphrase Extraction. Information Retrieval — INRT, 34-99.

[13] Word Net, A Lexical Database for English, Princeton University, http://wordnet.princeton.edu/.

[14] Zhang, C., Wang, H., Liu, Y. Wu, D, Liao, Y. and Wang, B. 2008. Automatic Keyword Extraction from Documents Using Conditional Random Fields. In *Journal of Computational Information Systems*, 4, 3, 1169-1180.