Extracting Keywords from Meeting Documents

Caslon Chua
Swinburne University of Technology
Faculty of Science, Engineering and Technology
H39 PO Box 218, Hawthorn, Victoria 3122, Australia
cchua@swin.edu.au

Abstract

Information plays an important role for management to make sound decision in an organization. As organization keeps documents in electronic format, it appears that aside from the possible, well-organised 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 analyse meeting documents to extract keywords in preparation for indexing and clustering.

Keywords: information search, information retrieval, natural language processing.

1 Introduction

Meeting is a way in which staff, managers, and administrators come together and discuss important issues in an organization with the purpose of making decisions, recommendations and suggestions among others. It is a key feature of a strategic process, both as part of the annual strategic planning cycle, and at times when critical strategic incidents arise (Jarzabkowski & Seidl, 2007). Meeting discussions are recorded as minutes 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 maintain meeting documents in electronic format, searching for specific documents can be quite tedious. Studies found that users do not file information according to keywords, but according to the notions of the kind of work that they do and the type of information they are dealing with (Barreau & Nardi, 1995). Even with the conscious effort of storing these documents in chronologically named directories and filename, it does not always provide one with an idea of what was discussed and recorded in the meeting document.

This brings about the need for desktop search applications to locate specific documents. In general, search applications process document as a single entity. This means that a keyword search would lead the user to the document containing the keywords, but the users will have to refine their search within the document viewer to locate the exact position of the keyword within the document. The user may also need to browse through the document before realizing the usefulness of the retrieved document. For instance, using the Windows Desktop Search requires a two-step search as shown in Figure 1. Step 1 of the search identifies and lists documents that contain the query string. Step 2 of the search requires the user to refine the search within the document preview.

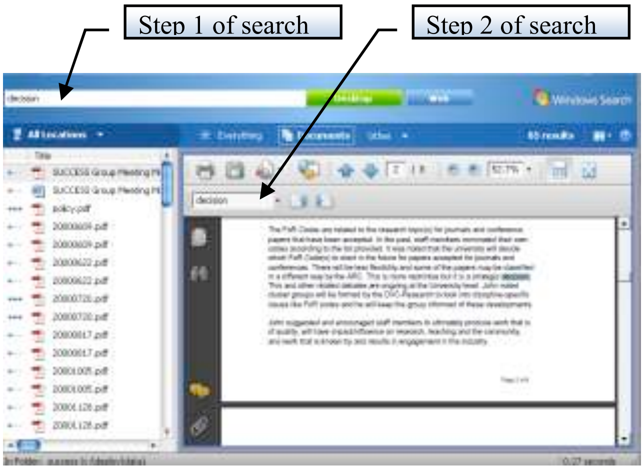

Figure 1: Document search using Windows Desktop Search

Another necessity in using existing desktop search application is the need to provide exact keyword or keywords in the query string. This means that keywords with different inflection will be missed during search. For example, words such as “decisions”, “decide” and “decided” will not be located if the query string is “decision”. In addition, if the document uses synonyms to imply a decision is made such as “resolution” or “chosen”, then that document will also not be located.

This study is on the development of an approach to parse and analyse meeting documents in order to extract keywords that can be used to enhance the search and

PCSC'14, March 6-8, 2014, University of Immaculate Conception, Davao City, Philippines.

retrieval process. This study first discusses different approaches in keyword extraction that can be used to represent actions within the meeting document. The study then focuses on extracting the keywords from each section of the meeting document, such as decision, recommendation, or suggestion. These keywords are then used to enhance document search when used in indexing and clustering. Finally, the study compares the keyword extraction technique against statistical approaches.

2 Related Literature

Keywords may serve as a dense summary for a document that lead to improved information retrieval, or an entry to a document collection (Hulth, 2003). In writing 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 (Turney, 2000). 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 (Turney, 2000).

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 (Gupta & Lehal, 2010).

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. (Luhn, 1957) proposed a statistical method towards supporting automatic encoding of documents for future information retrieval. (G. Salton, Yang, & Yu, 1975) 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. This is referred to as term frequency–inverse document frequency weight. Other term weighing approaches were later proposed by (Gerard Salton & Buckley, 1988). (Cohen, 1995) 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.

2.2 Linguistics Approaches

These approaches are based on the linguistics feature of the word, sentence and document. (Hulth, 2003) experimented on term selection approaches that include noun phrase (NP) chunks and terms matching any set of part-of-speech (POS) tag sequences. (Plas, Pallotta, Rajman, & Ghorbel., 2004) worked on automatic keyword extraction from spoken text using lexical resources. (Ercan & Cicekli, 2007) described a keyword extraction method that investigates the benefits of using lexical chain features.

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. (Frank, Paynter, Witten, Gutwin, & Nevill-Manning, 1999) described a procedure for keyphrase extraction based on the naive Bayes learning scheme. (Turney, 2000) described a hybrid genetic algorithm for keyphrase extraction.

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, Tian, Zhou, Ling, & Huang, 2006) described an algorithm that treats each document as a semantic network which holds both syntactic and statistical information.

3 The Prototype System

The prototype implemented in this study uses a rule-based approach in extracting keywords.

3.1 Document Analysis

In processing the documents, it is assumed that a partial structure exists in the minutes of the meeting documents as illustrated in Figure 2. The structure starts with the meeting name, day, time, venue and the members attending the meeting and finally followed by the meeting minutes. A decimal numbering system is used to number the topics in the document, topics are blocked in paragraphs and issues may run across several 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, topics are treated as sub-documents by the study. Each sub-document is analysed to allow extraction of keywords that can be used in indexing and clustering documents.

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 then mapped independently to the collection of sub-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, keywords are selected based its the frequency count within the document. For the linguistic approach, keywords are extracted if these words are tagged as verbs or nouns.

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 simple rule-based part-of-speech tagger. The tagger implementation assumes a noun-verb-noun sentence format is used in meeting documents. This implies that a verb is assumed to be in the middle of two nouns and is extracted to serve as the action of the sentence and subsequently the sub-document.

For example, given the following sub-document

The Council of Deans expressed no objection to promote/encourage the use of the indigenous language on campus. The members of the Deans' Forum will be requested to review the proposed written policy on this.

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

Noun	Verbs	Noun
Council of Deans	expressed	Objection
members of Deans' Forum	requested	Policy

Table 1. Part-of-speech keyword extraction

In this example, the tagger generates two actions for the sub-document based on the number of sentences, namely "express" and "request". "Council of Deans" and "members of Deans' Forum" are extracted as a single noun. Moreover when extracting nouns, various tolerance levels were used that allowed several numbers of prepositions and/or articles between two nouns. For example, if the text contains "Office of the Chancellor", it is also extracted as a single noun keyword.

3.3 Keyword Association

Aside from extracting keywords, these keywords are also expanded by mapping them to its synonyms. However, keywords composed of more than one word are not included in the expansion process. Synonym list from WordNet (WordNet) was used in this process. Figure 3 shows the keywords added to enable a non-exact keyword search for the term "decision".

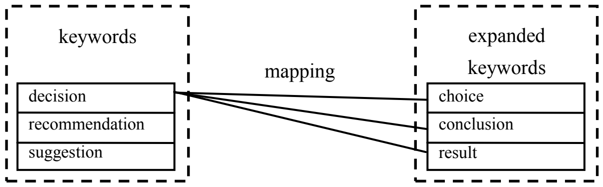

Figure 3. Expanded keywords list

4 System Test Results

In testing the prototype system, a collection of 103 meeting minutes from the Council of Deans' Meetings in a university covering 10 academic years are used.

4.1 Document Analysis

A rule-based tagger was implemented to analyse each document with 98% of the documents successfully parsed into sub-documents. Excluding the meeting information, each document has an average 289.38 words and containing average of 22.23 sub-documents. The prototype system ignored charts, figures and tables that may be present in the documents. In addition, bullet points used within a sub-document are interpreted as a paragraph.

4.2 Keyword Extraction

In comparing the keywords extracted based on the two approaches, the linguistic approach on average extracted the same number of verb keywords against the statistical approach based on document frequency. Table 2 summarises the average number of keywords extracted from each sub-document.

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*

Table 2. Keyword extraction count

In the statistical approach, the per sub-document statistics showed 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 Perceived Quality Results

Given that the document collection does not have keywords assigned to it, 13 participants were invited to evaluate the quality of the associated keyword mapped to the sub-document based on their perception. These participants have at least a bachelor's degree and are familiar with the conduct of meetings. The result showed that 35% of the participants' rated the association as fair, 36% as good or very good and 29% as poor or very poor. With fair perceived as an acceptable result, the result showed that shows that the implemented approach is promising. However, synonyms association were assessed to be generally fair. This may be attributed to the fact that the dictionary contains words that are outside the domain of an academic setting.

5 Conclusion

The prototype was able to demonstrate that an acceptable linguistic approach in keyword extraction can implemented in an unsupervised set up. Given that the manual assignment of high quality keywords is expensive, time-consuming, and error prone (Zhang et al., 2008), 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 utilised. 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 improving the tagger implementation, utilising an existing tagger such as the part-of-speech (POS) tagger developed by Stanford Natural Language Processing Group may be considered. In addition, noun selection can be improved to include negation such as "no" and "not". This will extract "no objection" instead of just "objection" from the above example in Section 3.2.2. In addition, "dis" and "un" prefixes can also be recognised. With the number of extractions based on the number of sentences in a sub-document, weight assignment can be used to select the main keyword to represent the sub-document. Thus the order of the sentence may be used to determine the weight assigned to keywords extracted from the sub-document. Other features of the extracted keywords will also be explored in future work. This can be based on various studies on text summarization extractive techniques described in (Gupta & Lehal, 2010). In addition, machine learning and hybrid approaches will also be considered.

In terms of application, these extracted and associated keywords can be used to enhance indexing and clustering used in search and retrieval. Furthermore, aside from using synonyms, the use of hypernyms and hyponyms in keyword association may also be considered. This will enable generalisation or specialisation when creating 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

Barreau, D., & Nardi, B. A. (1995). Finding and reminding: file organization from the desktop. *SIGCHI Bull., 27*(3), 39-43. doi: 10.1145/221296.221307

Cohen, J. D. (1995). Highlights: language- and domain-independent automatic indexing terms for abstracting. *J. Am. Soc. Inf. Sci., 46*(3), 162-174. doi: http://dx.doi.org/10.1002/(SICI)1097-4571(199504)46:3<162::AID-ASI2>3.0.CO;2-6

Ercan, G., & Cicekli, I. (2007). Using lexical chains for keyword extraction. *Inf. Process. Manage., 43*(6), 1705-1714. doi: 10.1016/j.ipm.2007.01.015

Frank, E., Paynter, G. W., Witten, I. H., Gutwin, C., & Nevill-Manning, C. G. (1999). *Domain-Specific Keyphrase Extraction*. Paper presented at the Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence.

Gupta, V., & Lehal, G. S. (2010). A Survey of Text Summarization Extractive Techniques *Journal of Emerging Technologies in Web Intelligence* (Vol. 2, pp. 258-268): Academy Publisher.

Huang, C., Tian, Y., Zhou, Z., Ling, C. X., & Huang, T. (2006). *Keyphrase Extraction Using Semantic Networks Structure Analysis*. Paper presented at the Proceedings of the Sixth International Conference on Data Mining.

Hulth, A. (2003). *Improved automatic keyword extraction given more linguistic knowledge*. Paper presented at the Proceedings of the 2003 conference on Empirical methods in natural language processing.

Jarzabkowski, P., & Seidl, D. (2007). The Importance of Meetings. 25. Retrieved from http://www.aimresearch.org/uploads/File/Publications/Executive%20Briefings%202/The_Importance_of_Meetings.pdf

Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM J. Res. Dev., 1*(4), 309-317. doi: 10.1147/rd.14.0309

Plas, L., Pallotta, V., Rajman, M., & Ghorbel., H. (2004). *Automatic keyword extraction from spoken text. a comparison of two lexical resources: the EDR and WordNet.* . Paper presented at the 4th International Language Resources and Evaluation, European Language Resource Association.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage., 24*(5), 513-523. doi: 10.1016/0306-4573(88)90021-0

Salton, G., Yang, C. S., & Yu, C. T. (1975). A Theory of Term Importance in Automatic Text Analysis. *Journal of the American Society for Information Science & Technology, 26*(1), 33-44.

Turney, P. D. (2000). Learning Algorithms for Keyphrase Extraction. *Inf. Retr., 2*(4), 303-336. doi: 10.1023/a:1009976227802

WordNet. (Ed.) A Lexical Database for English. Princeton University.

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