SYNONYM ANALYSIS OF E-JOURNALS FOR KEYWORD GENERATION

Richard Hussey

PhD, r.j.hussey@reading.ac.uk

ABSTRACT

The preliminary research into the area of the topic revealed there were no systems that employed similar methods, and there was in fact a complete absence of work into this as an area of research. The proposed system automatically analyses a given text document, comparing all the words and their synonyms, and produces suggestions for themes and keywords. The testing produced results that show an improvement over the base line and demonstrated that this initial work has potential for further uses in theme recognition.

1. INTRODUCTION

Current methods of keyword generation require data or information from outside the document being analysed, and this usually requires manual input to create and must be individually created for each document. Alternatively, they require more than one document to be analysed, and thus often the run time can be expensive both in terms of the complexity of the algorithm required and in terms of run time to process multiple documents.

The aim of the project is to develop a novel method of analysis that can, using synonyms, evaluate any arbitrary document and return to the user a useful and valid set of keywords that reflect the themes of the document.

The method explored in this paper is that of synonym analysis concerning the text of a document. Individual words are analysed for their synonyms and then ranked according to frequency, before the most common are presented as candidate keywords.

This paper will cover the background research into keyword generation and explore the lack of research into synonym analysis. Secondly, it will consider the proposed methodology for implementing the system to be able to analyse documents and, through experimentation, attempt to produce a system that can perform at least as well as manual keyword generation.

The following terms (Table 1) are in use in this paper, and their definitions given here for clarification.

Table 1 - Definitions

Term	Definition
Keyword	A word or short phrase comprised of three or four words used to identify the topic, theme, or subject of a document, or to classify the document. The collection of keywords for a document should be indicative of the major areas of interest within it.
Social bookmarking	Social bookmarking sites consist of publicly accessible lists of URLs that are generally tagged by the user who submitted them. Each user maintains their own list, and consequently the same URL can have different tags depending on which list is consulted.

Tag	A tag can encompass the notion of a keyword, but has additional uses in the web-community. Tags can be used to:
	• List meta-qualities of a document e.g. *NotWorthReading*
	  o Often used on social bookmarking sites
	• Reflect the author's state of mind at the time e.g. *bored*
	  o Often used on blog posts
	• Indicate future actions for the document e.g. *ToRead*
	  o Often used on social bookmarking sites
	• Group a resource that a particular group would be interested in e.g. *ThisIsMe*
	  o Often used on social bookmarking sites
	  o Also used as "Twitter hashtags" e.g. #ThisIsMe or #nptech (non-profit technology)

2. BACKGROUND

Previous research into automatic keyword generation has focused either on analysing a corpus of multiple documents and attempting to draw conclusions, or on taking a manual summary and attempting to extrapolate the reasoning to unseen cases.

2.1. Multiple documents

Multiple document approaches take a corpus and attempt to analyse relationships between the component elements to create methods for dealing with unseen elements.

Sood, et al proposes one such method, "TagAssist", in (1). Their system takes a corpus of blog posts and through comparisons with other blog posts attempts to supply a list of suitable tags for posts without any. The continually updated corpus incorporated new blog posts, to prevent the system from stagnating. The system first compresses the words of the blog into a series of stemmed words and then compares these to the stemmed words of tagged posts. Based on the tags found on other blogs, it then returns the top ten candidates as output. A panel of ten human judges, who ranked TagAssist second to the original tags, judged the outputs. They also found that in 51.15% of cases, the original manual tags were not appropriate.

Another example is given with song lyric keywords by Wei, et al in (2). This system used WordNet (3) to analyse relationships between words in a sentence, and to cluster them based on these links. Keywords are assumed to lie at the centre of these links, and the most similar ones across clusters are retained and compared with other songs.

In (4), the authors set out a system of directed graphs for linking keywords in non-obvious ways. This allows for the expression of statements such as "*relevance of keyword A to keyword B is independent of relevance of B to A*". The system set out by Joshi, et al, TermsNet, creates a framework to explore similarities of terms by placing them in context, along with additional terms found from a thesaurus. This is then used to weight links between the terms, and these weights can have different values depending on the starting term. The system produced keywords that ranked better than alternative systems on five out of the seven evaluation metrics that the authors proposed.

2.2. Manual summaries

Tuning via manual summaries has been used by a range of approaches to attempt to replicate the process by which a human can identify the themes of a document and reduce the text down to a shorter summary. These methods usually involve taking a corpus of texts for each of which a human summary has been created, and apply whatever method the authors have proposed to draw relationships between the original text and the summary. These relationships are then applied to a test set to see if they produce useful and usable summaries.

Li, et al (5) propose a system for classifying help desk problems by reducing the incoming e-mails, which vary wildly in their description of the problem, and often contain redundant or duplicated data, as well as being more loquacious than the summary. The manual summaries, created by the help desk engineers, are concise, precise, consistent, and formed from uniform expressions. The authors propose a system called a Stochastic Keyword Generator, which (during training) removes stop-words from a given text and then checks the resultant word list against the manual summary to see which of the words appears in both. By doing this between all of the text and summaries in the training corpus, it builds up a probability of any given keyword appearing in the summary, given the parent text. This method benefits from the ability to include words in the summary which are not in the parent text, but which have a high probability of occurring given the other words in the text (Figure 1).

In (6), Goldstein, et al set out a system for creating summaries based upon assessing every sentence of the document and calculating a ranking for its inclusion in a summary. The authors made use of corpora of documents for which assessor-ranked summary sentences already existed, and attempted to train the system to produce similar or identical sentences.

Figure 1 - Example of SKG (5)

3. IMPLEMENTATION AND METHODOLOGY

The implementation language for the system was chosen to be C#, for familiarity and ease of rapid prototyping.

The basis of this work is the examination of a document with reference to its synonyms and therefore the main bulk of the coding of the prototype system related to this and the associated thesaurus file.

The thesaurus (7) data representation was initially formatted into a symbol delineation scheme where different punctuation marks separated different parts of the thesaurus entries. Commas were used to separate the base word from its synonyms, and a hash was used to separate the synonyms from each other. This file was loaded into the program, and then stored in C#'s inbuilt dictionary data structure.

The dictionary structure consists of a Key and a Value, both of which can be of any data type. The Key serves as a unique index, and can therefore be easily utilised as an access to any Value. The Value is non-unique, and cannot be searched without enumerating through all entries (to ensure that none are missed). The base word was used as the Key, and the synonyms were stored as an array in the Value.

Data was loaded from the file, and used to create a list of unique words, which were then stemmed to remove plurals, derivations, etc. The resultant corpus of stemmed unique words was compared to the Key field of entries in the thesaurus, producing a list of the most common synonyms from the associated Value fields, and presented as a list ordered by decreasing frequency. The top ranked synonym was then returned as a keyword.

Additionally, these unique words from the file were run through the thesaurus from the other direction, and the results of this reanalysed for their synonyms. The arrays of synonyms (the Value of the dictionary) were searched for an instance of the word, and then the highest-ranking base word (the Key of the dictionary) was returned as a second keyword.

These two keywords are then presented as, generally, the same theme expressed in two different words. This is shown in Figure 2.

4. RESULTS

The system tested a number of papers taken from a collection of online e-journals, obtained from (8). There were five e-journals in this collection, each on a different topic and they were analysed separately. The topics were: Business Research Methods, E-Government, E-Learning, Information Systems Evaluation, and Knowledge Management.

For each of these e-journals, along with the paper itself, the authors supply keywords that they believe represent the paper. Therefore, the baseline metric for the system was to compare the keywords generated to those that the author supplied.

The results for each of the E-Journal are shown given below (Table 2).

Table 2 - Results of study

E-Journal	Papers	Matching	Percentage
EJBRM	72	17	23.67%
EJEG	105	14	13.37%
EJEL	112	63	56.37%
EJISE	91	25	27.50%
EJKM	116	32	27.60%
Average			29.70%

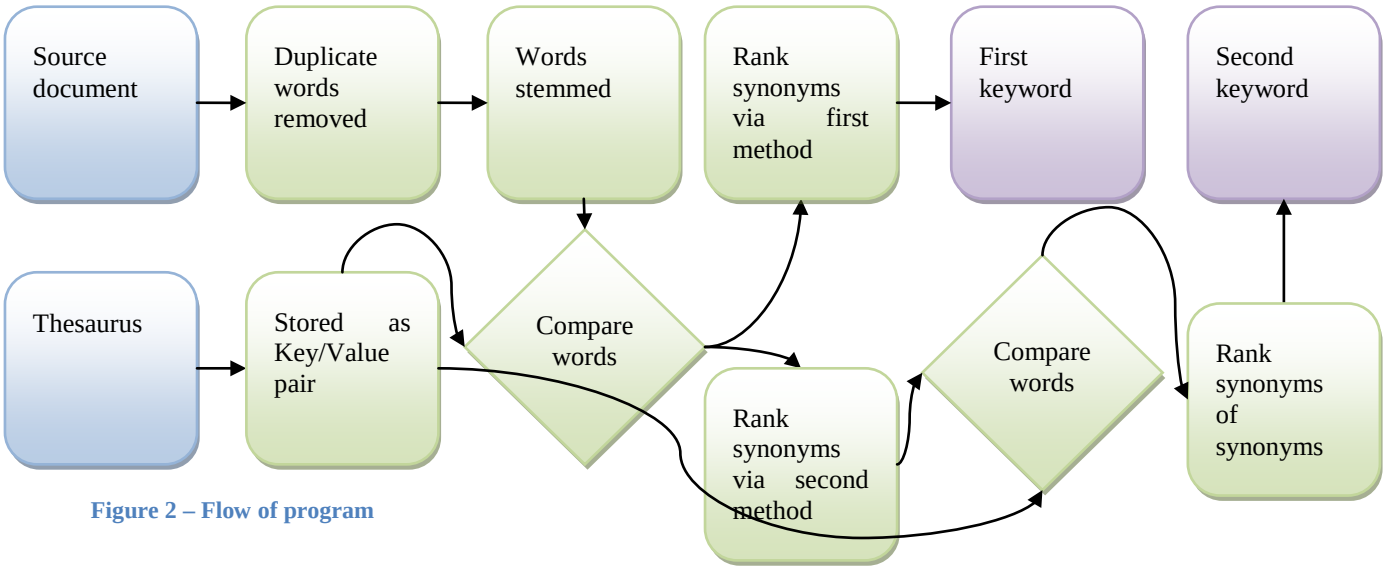

Figure 2 – Flow of program

5. DISCUSSION AND FURTHER WORK

The above results show that the initial version of the keyword generator produced keywords that matched manually assigned keywords on between 13% and 57% of the papers. There are a number of explanations for why the results differ as they do, and as to the low figures.

First, the words used in papers on these subjects are likely to contain jargon (subject-domain specific words) which are unlikely to occur in the thesaurus and, therefore, unrepresented in the program output.

Second, the keywords submitted by the authors can reflect their particular interpretation of the themes of a document (or what they believe the themes should have been), and those created by the system display an unbiased analysis. Additionally, the author's keywords can be included as a form of aide memoire, to remind them of areas of the paper to touch upon or to remind them about what it was they wrote.

Third, and linked to the second, the keywords submitted by the authors can often be more easily categorised as tags rather than keywords. Such examples include the 'keyword' "University of Birmingham", attached to a paper by an author who worked at that university. While this is a, potentially, valid tag, as a keyword it does not indicate a topic or theme to which the document holds: unless the document is actually about the University of Birmingham.

Fourth, this first prototype of the system outlined in (9) only identifies the most common theme of the document and thus can only reflect a single aspect of the document. This limits the accuracy of the keyword generation, as the manual keywords do not necessarily conform to the most common theme. Further work intends to extend the range of themes identified, and thus the number of keywords returned.

Fifth, the keywords returned are, by their nature as synonyms of themes, not always the same word for the same idea that the author has used. For example, if the author used "learning" and the system suggests "education", then the automatic comparison would reject it as unsuccessful.

Sixth, the thesaurus in use is the 1911 edition, as that is freely accessible at (7). This, therefore, does not contain the past one hundred years of new words, new word associations, nor does it remove disused words and associations.

Seventh, the current method of assigning counts to synonyms does not take any context of the words into account, and assigns equal increased to all synonyms, regardless of their appropriateness for the situation

These explanations provide a framework for carrying the project forwards. While not all of them are indicative of shortcomings on the part of the project, those that are form the basis of future work. The next stage of the project will look into recognising multiple themes of documents, and returning separate keywords for each of these. Later stages will examine the viability of reducing the effect of the other issues.

The explanations that affect the evaluation metric are ones that were outside of the scope of the project to date, but it is expected they will continue to have an impact on future stages of the project. As the current evaluation metric is based on the author submitted keywords, the success is based upon the accuracy of the manual tagging. In the course of this project, few corpora were discovered with a measurably useful keyword selection (which compares to the findings of (1), who found less than 50% of the original tags in their corpus were deemed appropriate by judges).

If a corpus could be identified that had appropriate keywords for its entries, or a team of people trained to produce relevant keywords for a training corpus, then this would have two effects:

1. To allow for better and more suitable comparisons of research methods in this area, regardless of who is undertaking the research.
2. To improve the accuracy of the evaluation of the reported systems currently using other methods (including the system presented in this paper).

6. CONCLUSION

The synonym analysis system developed by this project showed adequate results compared the base line, given the stated reasons for their low values. It showed that the initial concept of the project provides a sound basis from which to expand the work to improve the accuracy.

The identified issues presented an outline for future work, and a set of qualitative metrics for the evaluation of the results of the extensions.

The implementation demonstrated the use of the dictionary data structure as a powerful means of organising the data in such a way as to have it easily and quickly accessible to the program. It provides a platform on which to expand the project.

Acknowledgements: Thanks go to Professor Shirley Williams and Dr William Browne for their assistance as project supervisors throughout these initial stages of the project.

Thanks also go to the members of Odinlab for providing a sounding board for ideas during the project.

7. REFERENCES AND BIBLIOGRAPHY

1. **Sood, Sanjay C, et al.** *TagAssist: Automatic Tag Suggestion for Blog Posts.* Northwestern University. Evanston, IL : s.n., 2007. http://www.icwsm.org/papers/2--Sood-Owsley-Hammond-Birnbaum.pdf.

2. **Wei, Bin, Zhang, Chengliang and Ogihara, Mitsunori.** *Keyword Generation for Lyrics.* Comp. Sci. Dept., U. Rochester. s.l. : Austrian Computer Society (OCG), 2007. http://ismir2007.ismir.net/proceedings/ismir2007_p121_wei.pdf.

3. **Miller, George A., et al.** WordNet. *Princeton University.* [Online] March 2005. [Cited: 17 July 2009.] http://WordNet.princeton.edu.

4. *Keyword Generation for Search Engine Advertising.* **Joshi, Amruta and Motwani, Rejeev.** 2006. IEEE International Conference on Data Mining.

5. *Text Classification Using Stochastic Keyword Generation.* **Li, Cong, Wen, Ji-Rong and Li, Hang.** Washington DC : s.n., 2003. Twentieth Internation Conference on Machine Learning (ICML). pp. 464-471. https://www.aaai.org/Papers/ICML/2003/ICML03-062.pdf.

6. **Goldstein, Jade, et al.** *Summarising Text Documents: Sentence Selection and Evaluation Metrics.* Language Technologies Insitute, Carnegie Mellon University. Pittsburgh : ACM, 1999. pp. 121-128.

7. **Roget, Peter Mark.** *Roget's Thesaurus of English Words and Phrases.* 2004. http://www.gutenberg.org/etext/10681.

8. Academics Conferences International. *ACI E-Journals.* [Online] [Cited: 19 August 2009.] http://academic-conferences.org/ejournals.htm.

9. **Hussey, Richard.** *Synonym analysis of text documents for theme recognition and keyword generation.* School of Systems Engineering, University of Reading. Reading : s.n., 2009. p. 42, Transfer Report.

10. **Lin, Chin-Yew and Hovy, Eduard.** *The Automated Acquisition of Topic Signatures for Text Summarisation.* Information Sciences Institute, University of Southern California. Marina del Rey, CA : s.n., 2000. pp. 495-501. http://acl.ldc.upenn.edu/C/C00/C00-1072.pdf.