Discovering Encyclopedic Structure and Topics in Text

Laura A. Mather
Britannica.com, Inc.
3252 Holiday Ct.
La Jolla, CA 92037
mather@atg.eb.com

Jarrod Note
Britannica.com, Inc.
3252 Holiday Ct.
La Jolla, CA 92037
jnote@atg.eb.com

ABSTRACT
This paper presents a text-mining algorithm trained specifically for the discovery of encyclopedic structure and topics. Built on previous work on identifying subtopic regions and topics, this simpler algorithm performs two important tasks. First, the algorithm identifies subtopic regions within the text. This structure identification portion is trained and evaluated using encyclopedic table of contents headings as the subtopic boundaries. Second, the algorithm selects noun phrases to represent topics and subtopics both from the entire document and from subtopic regions. This topic identification portion of the algorithm is evaluated using the words in table of contents headings as topics. The results of the evaluations show that the algorithm achieves excellent recall and mid-level precision.

Keywords
Text mining, topic identification, subtopic region discovery

1. INTRODUCTION AND MOTIVATION
The advent of the World Wide Web has created a vast repository of text that is ripe for data mining analysis. As noted by Hearst [5], examples of that analysis include computational linguistic approaches, which focus on discovering various rules that can be used in such tasks as part-of-speech tagging, word sense disambiguation, bilingual dictionary creation, etc. Other text mining tasks suggested by Hearst include the discovery of new medical hypotheses based on the combination and subsequent analysis of large databases of text from related medical fields or the discovery of new topics in the topic detection and tracking scenario. This paper presents a less ambitious, though equally applicable, text mining application: given a set of ASCII text documents, the idea is to discover rules and algorithms that identify encyclopedic structure and encyclopedic topics in text.

Merriam-Webster's Collegiate Dictionary [12] defines an encyclopedia as "a work that contains information on all branches of knowledge or treats comprehensively a particular branch of knowledge usually in articles arranged alphabetically often by subject." Most encyclopedias have been compiled over many (sometimes hundreds of) years and therefore provide a thorough coverage of historically significant subjects.

In the current technology-driven world, knowledge is growing rapidly. In addition, access to that knowledge is often easier than ever given that so much information is available electronically. It would be advantageous for an encyclopedia company to be able to take advantage of the electronic publication of knowledge and to incorporate that knowledge as quickly as possible into the more traditional subjects listed in the encyclopedia.

This paper presents a first step towards the automatic incorporation of new text into an encyclopedic framework. The purpose of the algorithm is to use current encyclopedia articles to discover encyclopedic structure and topics in ASCII text. The algorithm first identifies table-of-contents-like structure and then attempts to find topical noun phrases within the text. The techniques described in this paper encompass a broad range of disciplines including natural language processing, information retrieval, and computational linguistics.

The remainder of this paper is organized as follows. Related work, including structure identification and topic discovery, is described in the next section. Then, Section 3 gives an overview of the algorithms used in this work. Next, the experiments used to evaluate the methodology are described including the design of the experiments, the data used, and the experimental results. Finally, some conclusions and the areas for future research are outlined.

2. RELATED WORK
Because the algorithm employs both structure discovery and topic identification, a review of earlier work on subtopic structure identification and topic discovery contributes to understanding the principles behind the algorithm presented in Section 3.

2.1 Finding Subtopic Regions in Text
The identification of subtopic boundaries in text was proposed by Hearst and Plaunt [6] as a means for improving document retrieval. They hypothesized that by finding subtopic regions in text and calculating the similarity of a query to

these subtopic regions instead of to the entire document, they could eliminate the noise of the various subtopics within each document and improve the performance of the retrieval algorithm. Hearst later refined this algorithm into what is now known as TextTiling [4].

The TextTiling algorithm is relatively straightforward. The text is divided into windows of n sentences and each window is compared with its neighbor windows using the traditional vector space model with a weighted cosine similarity measure. The similarities between all of the windows are then smoothed using a weighted mean and boundaries between subtopic regions are placed where there is a relative minimum in the smoothed similarities with respect to the surrounding similarity scores. An evaluation of the TextTiling algorithm shows precision of approximately 66% and recall at approximately 61% on magazine articles.

Another related method for segmenting text by topics is Beeferman, et al.'s exponential model for determining the document boundaries in streamed text [1]. Given a text stream (many consecutive documents without document boundary markings) the algorithm uses an exponential model to extract features that signify document boundary changes. The features include topicality features derived from language models and cue phrase features. The difference between the Beeferman approach and the method presented in this paper is that the subtopic regions found using the exponential model are more broad than those found in this paper since the Beeferman algorithm looks for document boundaries while the approach outlined herein looks for subtopic regions within documents.

2.2 Identifying Salient Topics in Text

Several groups are researching algorithms for finding topical phrases within text. Lin and Hovy [9] describe a study that characterizes the positions of keywords within sentences and paragraphs. The study concludes that titles contain 50% of keywords and that almost two thirds of the total keywords can be found in a fairly small number of sentences.

The work presented in this paper is an extension of the work done by Boguraev and Kennedy [2, 3, 7, 8] which identifies salient topics in text. The algorithm used in their work is described as follows. Their algorithm first finds subtopic regions using a methodology similar to Hearst's TextTiling algorithm described above. Phrasal units in the document are then isolated using the LINGSOFT text analysis tool. Coreference among the phrasal unit is identified and a salience score is calculated. The counts that are used in the salience score include the occurrence of the phrasal unit in a sentence, in a subtopic region, as the subject of a sentence, and elsewhere. The highest scoring phrasal units are returned as the salient topics of the text.

Boguraev and Kennedy point out that topic identification research is similar to text summarization research in that both fields strive to find a compact method for describing the content of a piece of text. The difference between traditional summarization research and topic identification research is that summarization research usually attempts to quantify the meaning of a text by extracting a subset of the sentences of the text while topic identification work is based on identifying phrasal units. Nevertheless, the summarization field should not be overlooked as a related area. Mani and Maybury [11] give an overview of recent advances in the summarization field.

3. METHODOLOGY

To review, the algorithm described in this paper has two purposes. First, it identifies subtopic regions within text. Then, it finds the topics associated with the subtopic regions. To find these two entities, the algorithm uses three basic modules. The first module tokenizes the text, the second module identifies the subtopic regions, and the third module finds the topics within those subtopic regions.

3.1 Tokenization

The tokenization portion of the algorithm is straightforward. Given a file containing ASCII text, LinguistX [10] is used to identify all of the noun phrases, sentence boundaries, and paragraph boundaries in the text. Each noun phrase, sentence boundary, and paragraph boundary is tagged for use in the other modules of the algorithm. After the noun phrases and boundaries have been tagged, all of the stopwords in the text are removed, all words are converted to lower case, and all of the remaining words are stemmed using the traditional Porter stemming algorithm.

3.2 Finding Subtopic Regions

The second module in the algorithm finds the subtopic regions in the stoplisted, stemmed text. This module is based on the Hearst TextTiling algorithm described above. The following list describes the steps performed to compute the subtopic boundaries in a file of stoplisted, stemmed ASCII text which includes sentence and paragraph boundaries.

- A window of three sentences is used to create sequential three-sentence blocks of text.
- A term vector is created for each of these blocks of text by counting the number of times each stoplisted, stemmed word occurs within the block.
- Tf-idf weighting is applied to each vector and the cosine similarity score between consecutive blocks of text is computed.
- The cosine similarity scores are smoothed using a median filter.
- The difference between subsequent smoothed similarity scores are calculated to determine whether the similarity between sentence blocks is increasing or decreasing.
- These differences in similarity are normalized by dividing by the $l2$-norm of the differences in the similarity. This normalization makes it possible to use the same threshold value for the difference in similarity between subsequent text blocks to determine subtopic boundaries for every text.
- Finally, a threshold is used to determine where the subtopic boundaries should be placed.

Once the subtopic regions have been found, tags are added to the stoplisted, stemmed text that identify where the subtopic boundaries occur.

3.3 Identifying Topics and Subtopics

After the subtopic regions have been found, a score is computed for each noun phrase across the entire text and in each subtopic region. The score indicates the salience of that topic in that region. The following list describes the parameters that are used in the score. Note that each parameter is computed for each noun phrase, both over the entire text and in each subtopic region.

- The percentage of sentences in which the noun phrase occurs (s).
- The percentage of paragraphs in which the noun phrase occurs (p).
- The percentage of subtopic regions in which the noun phrase occurs (r). This part of the score is only valid when considering the salience of the noun phrase in the entire document.
- The number of noun phrases that are subsumed by this noun phrase (N). (For example, the noun phrase *nuclear ceramics and low level waste* subsumes the noun phrases *nuclear*, *ceramics*, *nuclear ceramics*, and *low-level waste*.

The score for each noun phrase is then,

$$S = (s + p + r) * N. \tag{1}$$

The scores, S, are sorted, and the noun phrases corresponding to the top three scores for the entire document are returned as the topics of the text. The noun phrases corresponding to the top score for each subregion are returned as the subtopics of the text. This algorithm simplifies the score used in Boguraev's and Kennedy's work [2] in that it does not use part-of-speech information, co-reference information, or some of the other parameters in the earlier work.

3.4 An Example

The following example illustrates the goal of the algorithm described in this section. Consider the first two sections of an article from the Encyclopædia Britannica on *Nuclear Ceramics* [13].

Nuclear Ceramics

In their nuclear-related functions, *ceramics* are of major importance. Since the beginning of nuclear power generation, oxide ceramics, based on the fissionable metals uranium and plutonium, have been made into highly reliable fuel pellets for both water-cooled and liquid-metal-cooled reactors. (*) *Ceramics* also can be employed to immobilize and store nuclear wastes. Although vitrification (glass formation) is a favoured approach for waste disposal, wastes can be processed with other ceramics into a synthetic rock, or synroc, or they can be mixed with cement powder to make hardened cements. (*) All these *nuclear applications* are extremely demanding. In addition to severe thermal and chemical driving forces, *nuclear ceramics* are continuously subjected to high radiation doses.(*)

This article describes properties and applications of ceramics as nuclear fuels and as *waste-disposal materials*. For discussion of the employment of glassy and *metallic materials* in nuclear waste disposal, see *materials science*: Materials for energy. (*) For the *production of metallic uranium and plutonium* and their *conversion to oxide form*, see *uranium processing*. (*) For *detailed description of nuclear reactors* and the nuclear fuel cycle, see nuclear reactor. (*)

Nuclear Fuel

Ceramic oxide fuels were introduced in the 1950s, following military applications of nuclear power. Urania (uranium dioxide, UO_2) and plutonia (plutonium dioxide, PuO_2) have unique features that qualify them for nuclear fuel applications. First, they are extremely refractory: for instance, the melting point of UO_2 is in excess of $2,800^\circ$ C ($5,100^\circ$ F). Second, the open crystal structure of oxide nuclear ceramics allows for retention of fission products, and their highly variable oxygen-to-metal ratio can shift to accommodate burnup. They therefore have excellent resistance to radiation damage. (The *crystal structure of urania* is illustrated in Figure 2B of the article ceramic composition and properties: Crystal structure.)

Other advantages of oxide nuclear fuels include inertness to many coolants, long burnup without swelling, and relatively low fabrication cost. One drawback is low thermal conductivity. This has prompted research on replacing the oxides with more conductive carbides or nitrides. Selected properties of oxide, carbide, and nitride nuclear fuels are compared in Table 1.

The fabrication of ceramic nuclear fuels traditionally follows a standard powder-pellet process. This involves comminution, granulation, pressing, and sintering at $1,700^\circ$ C ($3,100^\circ$ F) in a reducing atmosphere. The resulting microstructure consists of large, equiaxed grains (that is, with dimensions similar along all axes), with uniformly distributed spherical pores on the order of 2 to 5 micrometres (0.00008 to 0.0002 inch). The pores are intended to retain fission gas and to decrease swelling during burnup. (*) ...

The (*)'s in the preceding text indicate the places where subtopic regions were identified by the algorithm. Notice that in the first two paragraphs, many subtopic regions were identified. Because the initial paragraphs give an overview of the article, they introduce many subtopics. The second section was determined to contain only one subtopic region. As it begins to develop one of the introduced subtopics, it is more specific than the first section.

The topics found to span the entire text are *nuclear ceramics and ceramic materials*, *disposal of radioactive and nuclear waste* and *T.O.Ma.* (which is part of the bibliography). The subtopics found for the above text are shown in italics. Although the subtopics are not always accurate, they tend to give an overall view of the content of the text when considered as a set.

4. EXPERIMENTS

Experiments were conducted to evaluate the performance of the algorithm. The following is an overview of the experimental design, the data used in the experiment and the results of the experiments.

4.1 Experimental Design

An evaluation of the algorithm described in Section 3 examined the performance of the algorithm's techniques on a large set of documents. Earlier evaluations of similar techniques were conducted on relatively small document sets: Hearst uses 13 articles [4] and Boguraev, et al. use one article and one reference manual [2, 3]. Hearst's results show average precision between 0.64% and 0.66% and average recall between 0.58% and 0.61% for discovering the boundaries between subtopic regions. Boguraev et al. do not give any indication as to precision and recall values, but the example shown demonstrates good performance in finding topical phrases in text.

Encyclopedia articles were used from the Encyclopædia Britannica with the purpose of discovering encyclopedic structure and encyclopedic topics in text. Encyclopedia articles comprise a realistic training set in that each article is devoted to a particular subject and are likely to be written by a unique author. The articles chosen ranged in subject from *Aesthetics* to *Paleo-Siberian Languages* to *Turtle*, indicating that the style of the articles is probably very heterogeneous. In addition, the 479 encyclopedia articles used in the experiments for this paper had a total of 614 unique authors (there can be more than one author per article). A training set was used to determine the value of the threshold in the TextTiling-like algorithm. Recall that this threshold determines the value of the similarity between two windows that indicates a change in subtopic. When the similarity between two sentence windows is below this threshold, a change in subtopic is indicated in the text.

After performing the training experiments and selecting a threshold for the subtopic identification algorithm, a test set was used to evaluate the performance of the overall algorithm. Two aspects of the overall algorithm were evaluated separately. First, the placement of the change in subtopic tags was compared with the placement of table of contents headings in the text. Finding a large data set of text documents with hand-marked change of topic identifiers is difficult. Given a set of encyclopedia articles, it was thought that the table of contents headings and tags indicating significant content mention are a reasonable first pass at human annotated changes in topic. (For simplicity, the remainder of this paper refers to the union of these two sets of words as "table of contents headings.") It should be noted that the authors believe that there are actually many more subtopic shifts in encyclopedia articles than those given by the table of contents headings. It is expected that any algorithm identifying changes in subtopic should at least identify table of contents headings as places where subtopics shift. In addition, the identification of more subtopic shifts than just the table of contents headings is reasonable, though it is difficult to say how many more subtopic shifts should be considered acceptable. Therefore, the recall of the algorithm, instead of the precision, was optimized for this experiment.

The second evaluation performed compared the noun phrases identified as topics and subtopics for the text with the words in the table of contents headings of the encyclopedia articles. To do this, stoplisting and stemming were performed on both the noun phrases from the text and the words in the table of contents. Then, every word in the identified noun phrases was compared with all of the words in the table of contents headings and those that occurred in both sets were counted as correct. Again, this is not an entirely accurate method for evaluation. Take, for example, this article. The table of contents headings would be *Discovering Encyclopedic Structure and Topics in Text*, *Introduction and Motivation*, *Related Work*, *Methodology*, *Experiments*, and *Conclusions and Future Work*. Only the title gives any information about the content of this document. In examining the encyclopedia articles it was discovered that while there are some table of contents headings that are not meaningful, most of the words in the table of contents headings contain some content information.

In addition to the issue of table of contents headings not containing content-full words, there is the issue of the number of subtopic regions identified in the first part of the algorithm. Because, as stated above, the algorithm is tuned to produce more subtopic regions than table of contents headings, and the noun phrase selection algorithm selects one noun phrase per subtopic region, there will also be many more noun phrases identified as subtopics than there are in the table of contents headings. Again, the authors believe that there are more subtopics in an encyclopedia article than those listed in the index. Therefore, the goal of this work is to optimize recall on the noun phrases selected as topics.

For these experiments 479 articles were randomly selected from the Encyclopædia Britannica. The following table gives some statistics about the articles used in this experiment.

Table 1: Data Set Statistics

	training set	*testing set*
number of documents	358	121
mean size of files (bytes)	51,934	40,004

Although there was a substantial difference in the average size of the files in each data set, it is believed that this did not influence the results of the experiment since the performance of the algorithm seems to vary on only extremely small files (5k bytes or less).

4.2 Experimental Results

The experiments tested two aspects of the algorithm described above: the performance of the subtopic region identifier and the performance of the topic identifier. The following table shows the performance of the algorithm in identifying subtopic regions in the test documents.

Table 2: Performance of Finding Subtopic Regions

average precision	0.4296
average recall	0.9826

In the above table precision is defined as

$$P = \frac{C_{str}}{T_{str}}, \tag{2}$$

where C_{str} is the number of correct subtopic region tags produced by the algorithm and T_{str} is the total number of subtopic regions produced by the algorithm. Note that a "correct" subtopic region tag is a tag that is placed within two sentences of a table of contents heading.

Similarly, recall is defined as

$$R = \frac{C_{str}}{TOC_{str}}, \tag{3}$$

where C_{str} is the number of correct subtopic region tags produced by the algorithm and TOC_{str} is the number of true subtopic regions (the number of table of contents boundaries).

Table 2 shows that although the average precision of the algorithm in identifying subtopic regions is fairly low, the average recall is very high. In essence, this means that the algorithm was able to find almost all of the table of contents heading breaks correctly and identified approximately 2 subtopic regions for every table of contents heading. This is reasonable since, on average, there is probably more than one subtopic in each table of contents region.

Next, the performance of the topic identification portion of the algorithm was tested. Recall from above that the words in the noun phrases identified as topic and subtopic words by the algorithm were compared to the words used in the table of contents headings. The following table shows the performance of this part of the algorithm.

Table 3: Perf. of Identifying Topics and Subtopics

average precision	0.2616
average recall	0.7621

In this case, precision is defined as

$$P = \frac{C_{tocw}}{T_{tw}}, \tag{4}$$

where C_{tocw} is the number of topic and subtopic words that matched words in the table of contents headings and T_{tw} is the total number of words found in the topic identification portion of the algorithm. Recall, on the other hand, is defined as

$$R = \frac{C_{tocw}}{T_{tocw}}, \tag{5}$$

where T_{tocw} is the total number of words in the table of contents.

Table 3 shows that the precision of the topic identification algorithm is fairly low. The poor precision performance is caused by the fact that for each table of contents heading there were approximately two subtopic regions found and that each of these subtopic regions was assigned one or more noun phrases (in the case of a tie, all noun phrases with the top score were returned as the topic or subtopic). The recall on this task is lower than expected. It is believed that, as part of the writing style in encyclopedia articles, different words are often used in the table of contents headings than are used in the text, which would result in low recall for this task. For example, an article on *Feline Leukemia* only uses the word "feline" once. All other references to felines in the article use the word "cat." This change in words between use in headings and use in article text negatively impacts performance scores. In addition, as stated above, the table of contents headings are often not informative about the content of the text.

5. CONCLUSIONS AND FUTURE WORK

This work is merely a first step in mining the information available in electronic databases such as encyclopedias, textbooks, reference manuals, and other large data sets. The algorithm mines the structure of the encyclopedia articles to determine appropriate subtopic regions based on the table of contents heading structure created by the articles' authors. It then uses these subtopic regions to identify topics and subtopics of text (in this case, other encyclopedia articles). It is believed that this type of work will someday facilitate the automatic incorporation of disparate electronic, textual databases into an encyclopedic format.

Although it is not possible to compare this work to other experiments, since no other research has used test sets of this magnitude, the results of this study are positive. Recall, which was to be optimized for this work, was very good in the subtopic region identification task and fairly good in the topic identification task. It may be possible to improve recall for the topic identification task by synonyms as correctly identified topics. In the case where "feline" is used in the heading and "cat" is predominantly used in the text, "cat" would not be counted as an incorrect topic.

Precision for both tasks could be improved by refining the algorithm described in Section 3. For example, it may help to add part-of-speech information to the score that determines the topics and subtopics. Also, the use of co-reference among the noun phrases would probably help. In addition, the use of various genre that contains more descriptive headings may improve precision.

Another interesting extension of this work would be to combine position information, as done by Lin and Hovy [9], with the information used by the subtopic region identification portion of this algorithm. Because many texts give overviews of their content at the beginning and/or end of the text, it may be possible to take advantage of this information when identifying structure, and possibly topics, of text.

Additional training may also help the algorithm. For example, it may be possible to determine coefficients for s, p, and r in the noun phrase scoring equation in order to optimize the noun phrases that are returned by the topic phrase identification part of this algorithm. Another example of the use of additional training would be to use several sets of training and test data instead of just one set.

Although this experiment is larger than any presented thus far, approximately 500 total documents is still relatively small in scale. In the future, further evaluation of the algorithm would benefit from running experiments on even larger test sets than those presented here. Also, it may be possible to find other electronic text genre that are amenable to the structure and topic analysis performed on encyclopedic texts for this work.

Acknowledgments The authors would like to thank Amy Steier for her insight into the potential applications of the algorithm described herein. In addition, comments from Johanna Jones, Mai Nguyen, Amy Steier and the anonymous reviewers significantly improved this paper. Finally, Alfred Inselberg's Parallel Coordinates research was instrumental in determining the optimal parameters for the subregion identification algorithm.

6. REFERENCES

[1] Doug Beeferman, Adam Berger, and John Lafferty. Statistical models for text segmentation. *Machine Learning*, 34(1), February 1999.

[2] Branimir Boguraev and Christopher Kennedy. *Salience-based content characterisation of text documents*, chapter 9. MIT Press, Massachusetts, 1999.

[3] Branimir Boguraev, Christopher Kennedy, and Sascha Brawer. An architecture for content analysis of documents and its use in information and knowledge management tasks. *SIGCHI Bulletin*, 30(2), 1998.

[4] Marti A. Hearst. Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd Meeting of the Association for Computational Linguistics*, 1994.

[5] Marti A. Hearst. Untangling text data mining. In *Proceedings of the 37th Meeting of the Association for Computational Linguistics*, 1999.

[6] Marti A. Hearst and Christian Plaunt. Sub-topic structuring for full-length document access. In *Proceedings of the 16th Annual ACM SIGIR*, 1993.

[7] Christopher Kennedy and Branimir Boguraev. Anaphora for everyone: Pronominal anaphora resolution without a parser. In *Proceedings of the 16th International Conference on Computational Linguistics*, 1996.

[8] Christopher Kennedy and Branimir Boguraev. Anaphora in a wider context: Tracking discourse referents. In *Proceedings of the 12th European Conference on Artificial Intelligence*, 1996.

[9] Chin-Yew Lin and Eduard Hovy. Identifying topics by position. In *Proceedings of the 5th Conference on Applied Natural Language Processing*. ANLP, 1997.

[10] LinguistX, distributed by XEROX-PARC, May 2000.

[11] Inderjeet Mani and Mark T. Maybury, editors. *Advances in Automatic Text Summarization*. MIT, Massachusetts, 1999.

[12] definition from Merriam-Webster's Online Collegiate Dictionary, http://www.m-w.com, May 2000.

[13] "Nuclear Ceramics", Britannica Online, http://www.eb.com:180/cgi-bin/g?docf=macro/5009/64.html, May 2000.