Agenda: A tool for agenda setting research

Yeong Haur Kok, Angela Goh a,∗ and Duncan Holaday b

a Division of Software Systems, School of Applied Science, Blk N4-2B-50, Nanyang Technological University, Singapore 639798, Singapore
b Division of Communication Research, School of Communication Studies, Nanyang Technological University, Singapore 639798, Singapore

Abstract. The paper describes the development of a software tool named Agenda to aid researchers in studying the correlation between the media and public agenda. Agenda processes news articles and forms clusters of topically coherent groupings. The use of clustering algorithms in document processing is therefore a key feature of the tool. In addition, the results are visualized to aid researchers in studying the data from different perspectives. Evaluation was carried out by comparing the results produced by Agenda with that obtained using an Intelligent Readers’ method involving human coders. Findings show that the results for the two methods are comparable, thereby highlighting the potential use of the tool.

1. Introduction

This paper stretches the range of overlap between the fields of information retrieval, with its roots in information science, and communication studies, a social science that focuses on the impact and effects of media. Through this attempt to apply computer-based methods of document comparison to agenda setting research, new areas for research in both fields have been identified.

‘Agenda-setting’ is a term used in communication studies to describe the effect the media has on the way readers think. The classic statement of the agenda setting hypothesis asserts that “the press may not be successful much of the time in telling people what to think, but it is stunningly successful in telling its readers what to think about” [2]. The term “agenda-setting” was first used in a study by M.E. McCombs and D.L. Shaw published in 1972 [6]. In the study, the researchers interviewed 100 undecided voters in Chapel Hill, North Carolina and asked them what issues they were most concerned about in the coming (1968) election. After determining the five issues the voters deemed most important, the researchers evaluated the media reports available to Chapel Hill for the content of their stories. McCombs and Shaw detected a high correlation between the types of stories that were covered most often and the voters’ concern for the same issues. Thus, the theory addresses how the media can influence where readers’ attention is focused. For example, the front page of a newspaper focuses out concerns and influences our thinking as to what is the crucial issue of the day. Much research into agenda-setting has since been undertaken resulting in hundreds of publications [7]. This topic is core to practically every textbook in communications theory [5].

The study of the effect of agenda setting involves the collection of two types of data: content of the media agenda and the public agenda. The two sets of issues are then compared, ranked and analyzed to investigate whether the effect of agenda setting has taken place. Presently, the process of determining the issues of the media agenda requires the use of professional readers. Their role is to read through

∗Corresponding author. E-mail: asesgoh@ntu.edu.sg.
all the news articles from a specified time period and note down the relevant data in a pre-determined template prepared by the researchers. Data collected include the prominence of the issues, the associated connotations, the frequency of reports and so on. At the same time, information on public agenda is obtained using a survey. The media and public agenda are compared and analyzed. The results have implications for the explication of the key concepts of “issue” and “readers”, and the underlying construct, “the media’s agenda” in agenda setting research.

In this paper, we focus on the aspect of the media agenda. Our work is motivated by the need for tools to reduce the effort required to determine media issues. The aim is to corroborate the results presented by professional readers and if deemed successful, the tool could be an alternative to such readers.

Currently, there are newspaper-publishers that maintain different types of electronic news over the Internet. For instance, the website, AsiaOne [13], created by Singapore Press Holdings Pte. Ltd. contains a number of on-line electronic newspapers such as LianHe Zaobao, Straits Times and Business Times. Readers can therefore access news articles not merely in traditional printed form but from various on-line newspapers. At the current rate information is produced and stored, more information is generated in a 24-hour period than a person can reasonably digest in years. With the support from browsers such as Netscape [10], people all over the world can access what appears to be a limitless source of knowledge. In order to cope with this flood of information, techniques for browsing, searching and visualization have appeared [1].

This paper describes the development of a tool called Agenda that will cluster news articles into relevant groups and provide researchers with various means to analyze the data. Users will have the flexibility to further investigate and study the relationships and trends of the media publications.

Following this introduction, a brief description of the system architecture of Agenda is provided. Details of the design and implementation of the main modules, which make up the system, are given in the following sections. An evaluation of results is given before final conclusions are drawn.

2. System architecture

Agenda is divided into four main modules. The first module, Article Search Module, culls relevant articles, which may be restricted using search queries. Downloading of news articles can be carried out prior to this to enable processing to be carried out locally without being connected on-line to any web site. User specifies the source database. Next, an Article Summary Module pre-processes the articles to extract the headlines, dates, word counts and so on. Useful functions to view, save and print the articles are also provided. This is followed by an Article Processing Module, which performs stopping and stemming and the generation of word frequency for each article. Stopping removes common function words from the articles, while stemming changes variations of words to their root forms. Finally, the Cluster Recognition & Representation Module performs clustering by grouping the articles into topically coherent clusters. Information such as descriptors, made up of the highest weightage common words found in the cluster, and the document set is produced for user evaluation. The results of clustering and all the associated information are available for user viewing. An outline of the process taken by Agenda is given in Fig. 1.

Agenda was developed using Microsoft Visual Basic 5.0 on a Pentium PC running Windows 95 Operating System. The compiled program is an executable application that can run on any 32-bit Windows platform.
Fig. 1. An overview of Agenda.
3. System functionality

In this section, details of the main modules of the system will be discussed. These are the Article search and summary, Document processing and clustering.

3.1. Article search and summary

The articles that are to be processed have to be available to Agenda. These articles may have been scanned in from newspapers. More likely however, such articles are downloaded from an on-line news source. The system is able to process Hypertext Markup Language (HTML) format or text format. The researcher may wish to perform clustering on a selected subset of articles, for example those related to a particular topic such as foreign affairs or financial news instead of working on the entire news archive. This module provides the interface with which the scope can be limited. This is achieved by specifying the types of articles to process via a search query. Accommodating a maximum of five keywords, the program searches for the articles in a pre-specified directory. To refine the queries, Boolean operators, AND and OR are available. In addition, the options of exact phrase and/or exact case are also available.

Following a possible restriction of the articles, an extraction of pertinent information such as headline, date of publication, word count is obtained. This information is presented to the user in a summary display. The summarised information is stored in a data structure necessary for further processing. The process is illustrated in Fig. 2.

As an option, any of the articles can be viewed, saved and printed. The summary includes the word count frequency, a tabulation of the unique words making up the article and their frequency of occurrence. The generation of this word frequency is described in the next section.

For articles in HTML form, viewing is available in two modes – in their native form using a simple built-in browser provided by Agenda or in text format. For the latter, the program will automatically convert the HTML documents to extract the relevant content without the accompanying tags.

3.2. Article processing

The Article Processing Module serves to build a word frequency list. Figure 3 shows the process through which such a list is built. The process of stopping and stemming are essential. The former removes common words that are unlikely to contribute to the ranking process. For example, *he, it, and this* are treated as stop words because they are too common to play a significant role as keywords. The stemming procedure translates words to their root form. For instance, the word *automobiles* is changed
to *automobile*. Both procedures refer to a stopword list and a stemming list provided by WordNet [9] to locate and process the words in the document. After the stopping and stemming processes, the remaining words are stored in user-defined structures together with frequency of occurrence. However, word frequency alone does not completely convey the true weightage or importance of the word in indicating the article's content. Thus, the significance of word frequencies is not absolute but relative to other factors such as the article's total word count and the frequency of other words. Therefore, an Inverse Document Frequency (IDF) which measures the goodness of particular terms and acts as a document discriminator is used. As these measures are used in clustering, it is crucial that they reflect the article correctly. The IDF is given by the following equation [12].

\[
IDF_1 = \log_2 \frac{\text{max } n}{n_i} + 1,
\]
where \( n_i \) is the total number of occurrences of word \( i \) in the collection. Max \( n \) is the maximum frequency of any word in the collection.

3.3. Clustering

The measures computed in the Article Processing Module are utilized to produce clusters. There are many clustering algorithms available [4,11]. Several algorithms were implemented, namely Similarity score, Group-average-linkage cosine measure and Buckshot [3]. Similarity score is the simplest of the three. It has the advantage of fast processing speed but a drawback of poorer cluster quality. Group-average linkage is akin to Similarity score with the exception of using cosine measure as the comparison criteria. Finally, Buckshot offers advantage in processing speed for large volumes of articles. No comparisons of the various methods will be presented in this paper except to report that the Group-average linkage cosine measure had significant advantage in terms of the quality of clusters produced. For the remainder of the paper, clustering will refer to the Group-average-linkage method. The results are presented to the user in three different perspectives described in Section 4.

4. Presentation of results

4.1. Listview display of cluster results

The listview display is a summary of the clustering results and is the first display presented to users after clustering has been completed. Each cluster is accompanied by a set of descriptors to indicate their content. Upon selection of any cluster, the set of articles making up the cluster and the common words they contain are displayed. Headlines and pathnames of the documents, as well as weightage of the common words are also provided. Any documents listed in the set can be viewed, either in its entirety or in its word frequency composition. From this display, user can choose to visualize the clusters or view a chart display to further analyze the results. Figure 4 shows a sample of the listview display.

As shown in Fig. 4, clusters formed are listed in two separate windows, one for multi-document clusters and one for single-document clusters. As the term implies, a single-document cluster consists of one document only. Such single-document clusters may arise from content incompatibility with other articles or other reasons related to its word composition and characteristics. Due to the nature of news articles, the presence of a large number of such single-document clusters can be expected. It is often the case in daily newspapers that reported issues are isolated incidents or events, such as robberies, accidents or newsworthy happenings.

4.2. Visual representation

The visual representation of the clusters enables a more natural way of studying the relationship of the clusters with one another. Furthermore, visualization offers a more user-friendly graphical interface for users to investigate the characteristics of the clusters.

Clusters are represented as circles, the size being a function of the number of common words contained within the cluster. Each circle is identified by a unique number and the size of its document set is shown in brackets below the number.

Pop-up menus are used to display a cluster’s properties. Information available includes pathname and headlines of the document set, the descriptors for the cluster, and its intersecting relationship with the other clusters. Figure 5 shows a popup menu in action in the visual display.
Fig. 4. Listview of cluster results.

Fig. 5. Cluster visualization with pop-up details.
Intersecting information between the clusters can be further investigated through another detailed display as shown in Fig. 6. This display is presented when any of the circles in Fig. 5 is left-clicked. Seemingly different clusters may be related to one another through similar issues mentioned or reported in both of them. The provision of this display helps to identify these links. The user can choose any of the other circles to compare with the selected cluster. Results of the comparison, namely the intersecting words between these two circles, are displayed at the bottom right panel of the form.

4.3. Timeline display

The purpose of this display is to plot the distribution of the clusters over time. This is accomplished by plotting an XY scatter graph of the word count of the articles against the dates of the articles appeared. This timeline display reveals temporal information while word count acts as an indication of the prominence of the article.

This facility is a particularly useful feature for investigating the spread or concentration of issues over time. For example, researchers may discover certain topics surfaced or were featured prominently during a particular period, hence revealing an issue emphasized by the media. The issues do not have to be concentrated in any particular “peaks” in time to yield useful analysis. Some issues may be discussed pervasively during a certain period but still continue to be mentioned, albeit with a lower frequency, in the ensuing time periods after the “peaks”.

A maximum number of six clusters are allowed at any one time on the chart. This user-definable limit is imposed to ensure the chart is not overloaded with clusters, resulting in an unclear presentation. Upon selection of the clusters to view, the chart is plotted and displayed. Figure 7 shows a sample of a timeline chart for selected clusters.

Copyright © 2000. All rights reserved.
Each cluster is assigned a unique symbol and hence, the distribution of articles in a cluster over time can be observed from the chart. These symbols are color-coded to facilitate clearer visualization. Date and word count labels of the documents can also be displayed. A Replot function is provided to redisplay a different set of clusters. For easy reference to the clusters’ content and information, a treeview is provided in the chart for easy navigation. Each entry in the hierarchical listing is represented by a short description, which can be expanded to reveal its document set.

5. System evaluation

This section focuses on the comparison between the results produced by Agenda and the Intelligent Reader Method. The news articles used for testing and analysis were taken from a Singapore newspaper, the Straits Times, for the period between 8th and 14th February 1996. These articles were obtained from the archives of Newslink and Reuters. A total of 227 articles were culled from daily issues during the test period. Instead of analyzing the entire newspaper, articles were taken from pages one to three, the Home section for local news and letters to the editor. No further search restriction was imposed on these articles. The length of these articles range from 57 to 1710 words with the average being 407 words.

5.1. The intelligent reader method

As explained in Section 1, agenda setting investigation involves the collection and classification of news reports, from which meaningful and indicative trends may be observed. This process primarily
requires the classification of news articles taken within a predetermined period of time into topically coherent groupings.

A method known as the “Intelligent Reader” coding was employed by the researchers [8]. As part of a longitudinal study of agenda setting in the 1996–97 Singapore elections, human coders were asked to read the above-mentioned sections of the Straits Times. The researchers provided a coding sheet containing specific categories. After reading each day’s paper, the coders provided a value on a six-point scale to each of the articles read. The values ranged from 0 (no attention at all) to 5, which signifies the category been an outstanding concern of the day. The overall emphasis was averaged for each week using double coding for reliability. This method of coding differs from the traditional procedure of identifying issues by title or keywords and measuring emphasis by column inches. Highlighting the shift of emphasis to the coder’s judgment, this method was dubbed “the intelligent reader method”. The process of this human coding method is illustrated in Fig. 8.

5.2. Comparing results of Agenda with human coders

A total of 31 multi-document clusters and 128 single-document clusters were produced by Agenda. Each cluster produced was grouped to one of the 17 categories used by the human coders. These categories were then ranked in order of prominence by summing the total number of documents of the clusters grouped under them. The ranking was then compared with that of the human coding method, which was based on the value of attention accumulated.

The categories ranking for the two approaches are listed in Table 1.

The table shows that there is a fairly high correlation between the two approaches in terms of measuring issue emphasis. The categories “government policies” and “election” featured prominently using both approaches. The lack of emphasis given to issues such as “poverty”, the “general economy”, “social class”, “media” and “foreign interference” were apparent in both rankings as well. This shows that the agenda program performs reasonably well in the identification of media emphasis.
### Table 1

Comparison of categories between human coders and Agenda

<table>
<thead>
<tr>
<th>Intelligent reader</th>
<th>Agenda</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Government Policies</td>
<td>• World Affairs</td>
</tr>
<tr>
<td>• World Affairs</td>
<td>• Personal Wealth and Materialistic concerns</td>
</tr>
<tr>
<td>• Election, voting and politics</td>
<td>• Election, voting and politics</td>
</tr>
<tr>
<td>• Housing &amp; Upgrading</td>
<td>• Government Policies</td>
</tr>
<tr>
<td>• Health</td>
<td>• Cost of living; People (tied)</td>
</tr>
<tr>
<td>• Education</td>
<td>• Housing &amp; Upgrading; Education (tied)</td>
</tr>
<tr>
<td>• Values</td>
<td>• Health; Values (tied)</td>
</tr>
<tr>
<td>• Ethnicity, race &amp; chauvinism</td>
<td>The following categories were listed low in the list.</td>
</tr>
<tr>
<td>• Cost of living; People: Workers’ rights/unions (tied)</td>
<td>Poverty and subsistence of living; General</td>
</tr>
<tr>
<td></td>
<td>economy/productivity; Ethnicity, race and</td>
</tr>
<tr>
<td></td>
<td>chauvinism; Social class, inequality; Workers’</td>
</tr>
<tr>
<td></td>
<td>rights/unions; Media issues</td>
</tr>
<tr>
<td>• Personal Wealth and Materialistic concerns</td>
<td>The following categories were listed low in the list.</td>
</tr>
</tbody>
</table>

The following categories were listed low in the list: Poverty and subsistence of living; General economy/productivity; Ethnicity, race and chauvinism; Social class, inequality; Workers’ rights/unions; Media issues

### 6. Conclusion

The development of a software tool to aid researchers in the investigation of Agenda Setting involved the ability of computers to cluster textual information into topically coherent groupings. Comparison of the results produced using Agenda and the Intelligent Reader method revealed potential usefulness of computer-based content analysis in Agenda Setting research. The use of clustering in aiding Agenda Setting analysis is but one of the many possible applications of document clustering. This demonstrates the potential benefits and feasibility of computer-based textual processing in real-world applications. Although the use of cluster analysis produced promising results in grouping news articles, the results also show that ultimately, numerical data is an imperfect reflection of reality. Without the use of natural language processing (NLP), there is a limitation to what statistics can tell us about the content of text. The cosine measure assumes a vector space such that two items \( x \) and \( y \) with only a single word occurrence each will always have a cosine measure of 1 when they have the same word and 0 otherwise. While mathematically correct, this is contrary to semantic reality. This implies that all words processed are independent of each other which is clearly false in any natural language and especially in semantically rich languages such as English. Hence, there is a possibility that two articles reporting on the same incident can be classified under two different clusters when the words used to describe the incident are different. The incorporation of an NLP engine would clearly strengthen Agenda’s ability to produce more meaningful clusters.

### References