Paper # DM08

A TEXT MINER ANALYSIS TO COMPARE INTERNET AND MEDLINE INFORMATION ABOUT ALLERGY MEDICATIONS

Chakib Battioui, University of Louisville, Louisville, KY

ABSTRACT

Recently, the internet has become the primary resource for consumers seeking medical information. Pharmaceutical companies are marketing directly to consumers. An important concern for many physicians is the quality of health information online. The purpose of this paper is to examine the issue of internet marketing using SAS text mining software. The specific objective is to examine information about allergy medications from Medline and from the internet.

The search term “Allergy Medications” was used in Medline through the Ovid search engine. A total of 115 papers were returned. A similar search was performed using the search engine Google. A total of 100 sites were selected. The two data sets were combined and analyzed using SAS Text Miner 9.1. The analysis clearly shows a difference between the information contained on the internet compared to that contained in medical journals. Most of the medical information online is about the type of allergy drugs and what those drugs are for, while that information is almost completely absent in Medline. On the other hand, most of the medical journals contain information about clinical studies and research. Only a little of this information was found online.

INTRODUCTION

Over the last few years, companies have been struggling with enormous mountains of unstructured, text data, such as e-mails, web pages, letters, reports, customer surveys, notes and medical records. There is too much information that needs to be read and analyzed. Text mining is used to evaluate this flood of electronic information and to avoid the overwhelming task of reading. It is the process of identifying concepts from unstructured text in document collections using methods that permit further analysis.

Today, the internet has become the primary resource for consumers seeking medical information. Pharmaceutical companies are marketing directly to consumers. An important concern for many physicians is the quality of health information online. The purpose of this paper is to examine the issue of internet marketing using SAS text mining software. The specific objective is to examine information about allergy medications from Medline and from the internet, and then to extract concepts using SAS text mining software.

Two primary document sources were used. The first, Ovid®, provides a mechanism for searching for Allergy Medications from the electronic database Medline [1], a database provided by the National Library of Medicine. It contains abstracts from thousands of medical journals. Ovid provides front-end access to Medline. The second, Google®, provides links to other websites on Allergy Medications, including many articles and reports as well as advertisements. The search term “Allergy Medications” was used in Medline through the Ovid search engine. A total of 115 papers were returned in the search. We performed a similar search using Google and “Allergy Medications”. The first 115 documents returned were used since most people examine the first sites on the list.

BACKGROUND

According to a report from a Pew Internet and American life Project, 77 million American adults said they go online to look for health or medical information [2]. Eight out of ten of those who have conducted health searches said they do so at least every few months. Many people who seek health information for themselves say that this information affects their choice of treatment. Since so many consumers are using this information, it becomes important to assess its quality. However, identifying relevant and valid information can be problematic. Misinformation can pose a real danger and can lead to harm. Many studies focus on evaluating the quality of medical information online and on developing criteria and methods for assessing the quality of health information on the Internet [3]. The major criteria are credibility, content, disclosure, links, design, interactivity and caveats.[4]

Quality is an issue because of the presence of bogus or false information on the internet. There is no governing agency to accredit medical information or to remove information that is false. Anyone with a computer and internet access can create a web page, and register it with search engines to make it easily accessible. On the other hand, publishing on Medline is not an easy matter. Medline is a medical database of the US National Library of Medicine. It catalogues more than 4000 medical journals published from many countries starting from 1966. It is a useful tool for identifying articles on issues in clinical medicine and related fields. The quality of content is one of the key factors in recommending a title for indexing. The decision whether or not to include a journal for Medline, is made by the Committee of the US National Library of Medicine. A journal must satisfy many conditions. For example the Journal of the American Medical Association (AMA) requires authors to meet the following basic criteria: material is original;

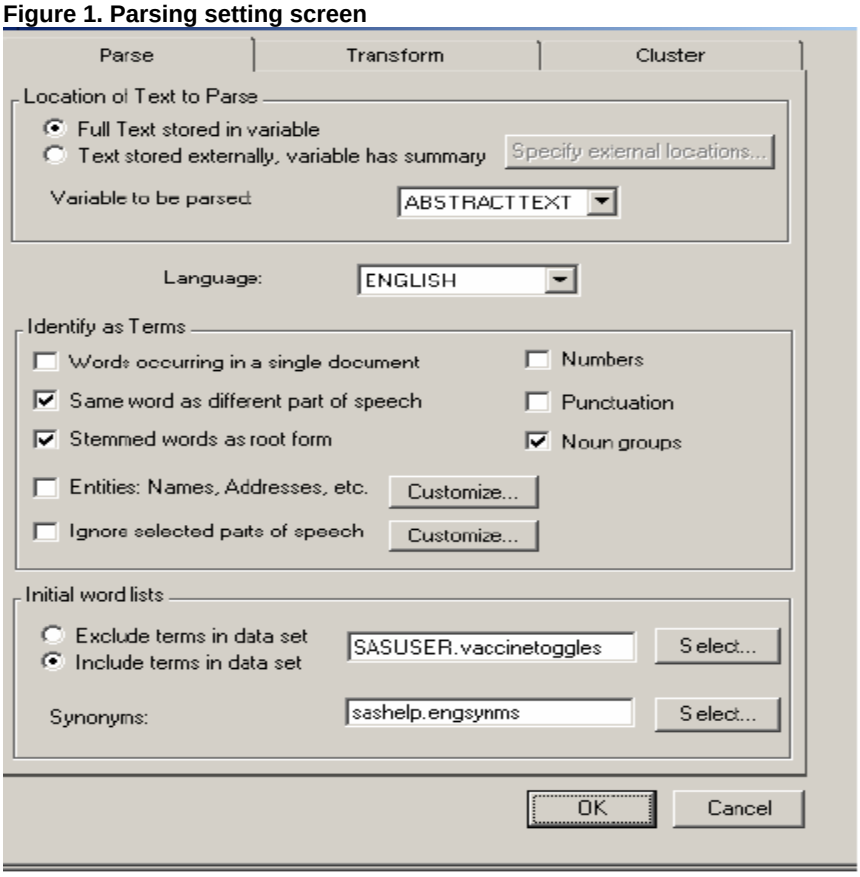
writing is clear; study methods are appropriate; the data are valid; conclusions are reasonable and supported by the data; information is important; and the topic has general medical interest. Only 8% of the papers received by the AMA get published every year [5].

METHOD

The search term “Allergy Medications” was used in the electronic database Medline through the Ovid search engine. A total of 115 abstracts were returned in the search. An Excel database was constructed containing those 115 citations with Authors, Source, Date, Title and Abstract as defined fields. Similarly, we performed a search online using Google. The search returned 317,000 matches. Of the top 300 sites, we selected the first listed 115 and entered them in the Excel spreadsheet as well. The internet (115 web sites) and Medline abstracts (115 abstracts) were concatenated into one dataset and clustered using Text Miner.

Text Miner has three setting screens allowing the user to choose from different options in order to cluster and analyze the unstructured text. The first setting screen is about parsing, It is given in figure 1.

Figure 1. Parsing setting screen



The purpose of text parsing is to group similar documents based on the terms used within the document. It is used to reduce the size of the documents to a manageable number. There is an option about the location of text to parse: either the text is stored in a SAS dataset, or there is a variable in the data set that points to the location of the documents.

The first box is unchecked allowing Text Miner to exclude words that only occur in one document. The second and third boxes are checked to consider a word to be different if it is used as a different part of speech, and to consider words with the same stem respectively. Numbers and punctuation are excluded for clustering text documents. However, Noun groups can be used for clustering.

Stop words are words that have no consideration in identifying documents. A standard stop list will remove words such as the, and, of. However, the user can add words to the stop list as needed.

The second setting screen is given in figure 2. It allows for the user to determine the method to work with the parsed matrix and to weight the value of each term in the documents. There are two main methods to reduce the parsed matrix to a manageable size: the singular value decomposition (SVD) and the roll up terms. The singular value

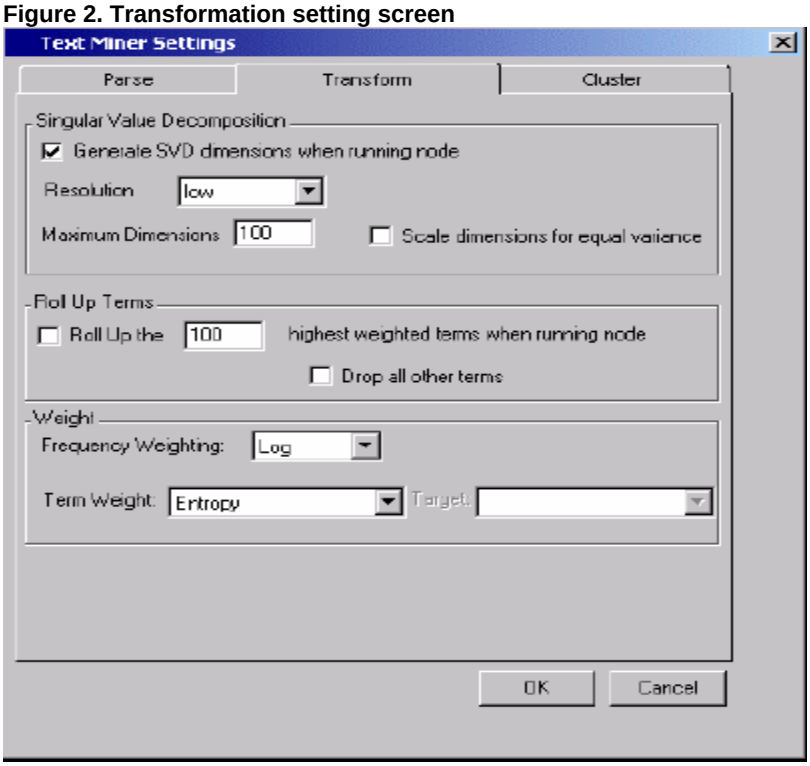
decomposition provides a wealth of information to the analyst, including orthogonal bases for the domain and range spaces. The SVD of a matrix A is:

$$A=U\Sigma V$$

where U is the matrix of term vectors, Σ is a diagonal matrix with singular values along the diagonal and V is the matrix of document vectors. The user is allowed to set the maximum size of the matrix (by default 100). Roll up Terms are by default, the 100 terms with the largest term weights. The term-document frequency matrix is just the number of roll up terms by the number of documents. No further dimensionality reduction occurs [6].

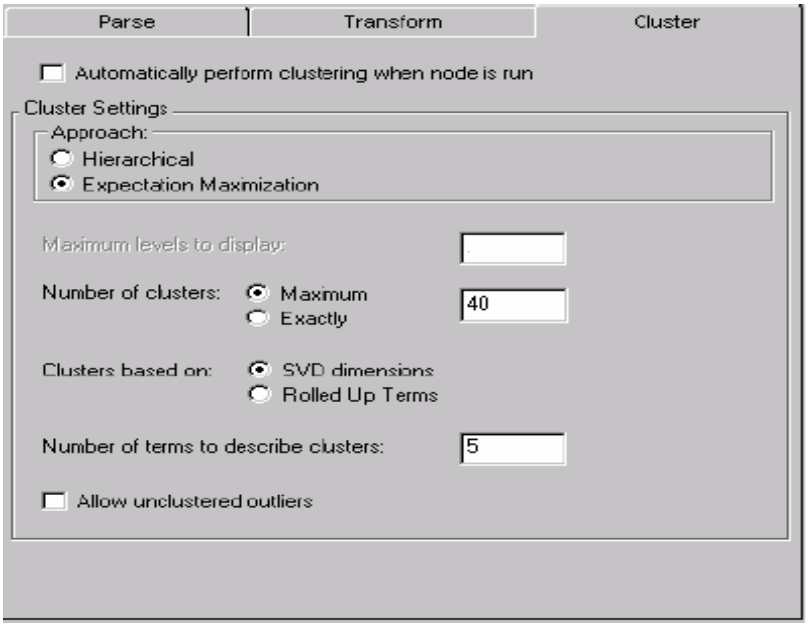
Term weightings are used to identify that some terms are more important than others. The discriminating terms will help separate documents. Entropy is the default weighting. Terms that appear more frequently will be weighted lower compared to terms that appear less frequently [7].

Figure 2. Transformation setting screen



The last setting screen is given in figure 3.

Figure 3. Clustering setting screen.



Clustering divides data into mutually exclusive groups based on distance or similarity measures. These groups need not have any real meaning or any value in describing the data, although a successful clustering will support meaningful results. Text Miner supports two clustering algorithms: Hierarchical and Expectation Maximization. The user can set the method of clustering, and also the number of clusters. The user can also specify the number of terms to be used to cluster. We changed the number of words from 5 (default) to 20 in order to label the clusters effectively. We used the defaults for the other features.

RESULTS

Text Miner returned the clusters contained in table 1.

Table 1. Cluster Contents

Cluster	Descriptive Terms	Frequency	Proportion of Documents
1	Evaluation, man, patient, objective, associate, efficacy, examine, study, association, conclusion, woman, compare, group, suggest, evaluate, safety, risk, data, stroke, significantly, outcome	64	0.309178744
2	Aggressive, observe, var, function, else, return, therapeutic, true, intervention, ad, airflow, chronic state, day-to-day, expire, netscape, random, secure, smooth muscle, bind, bottom, composit	6	0.0289855072
3	Allergic rhinitis symptoms, brief, clinical trials, cromolyn sodium, allergic rhinitis sufferers, difference, cromolyn, sodium, effort, active, ad, airflow, chronic state, much, daily, attention	4	0.0193236715
4	Allergies, doctor, food, air, news, anaphylaxis, contact, medicine, animal, back, hive, body, breathe, even, important, professional, help, itch, cause, dust, swell, copyright, follow, late	39	0.1884057971

We allowed a maximum of 40 clusters; a total of 4 were returned. The first column in table 1 shows the descriptive terms that were used to identify each cluster. The second and third columns show respectively the frequency and the proportion of documents of the descriptive terms on the combined text. We labeled every cluster based on the context of the descriptive terms. The labels are given in table 2.

Table 2. Cluster Labels

Cluster	Label
1	Clinical Trials

2	Therapeutic Intervention
3	Allergic rhinitis
4	Type of allergies

We used a chi-squared analysis to make a comparison between data sources. Results are given in table 3.

Table 3. Comparison between Medline and Internet information

Cluster	Label	Frequency	Web	Medline
1	Clinical Trials	64	7.81	92.19
2	Therapeutic intervention	6	66.67	33.33
3	Allergic rhinitis	4	50.00	50.00
4	Type of allergies	39	94.87	5.13

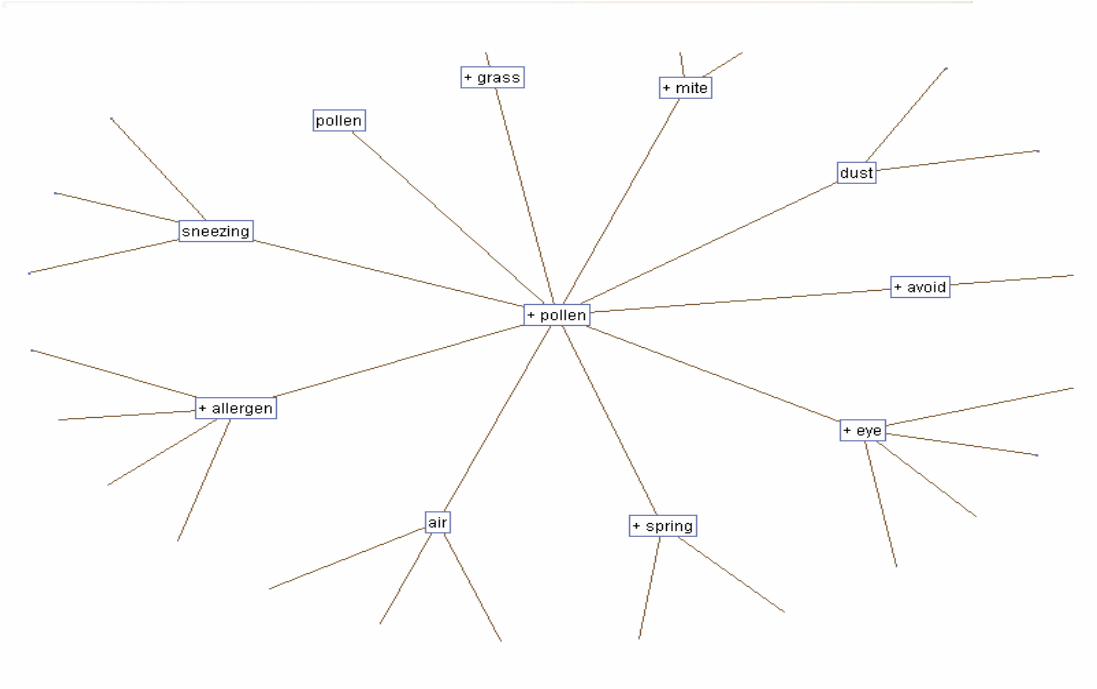
The results clearly show a difference between the information contained on the internet compared to that contained in medical journals. Most of the Google information online is about the type of allergy medication drugs and what those drugs are for, while we observe almost the absence of that kind of information on Medline. On the other hand, most of the medical journals are about clinical studies and research, very little of that information was found online.

RESULTS OF CONCEPT LINKS

Concept links are graphs consisting of word terms that are connected to each other. The nodes are labeled with descriptive text, representing the "concept". Concept links are an important means of knowledge representation because many people find them intuitive and easy to understand. Concept maps have been used in many fields including education, management, artificial intelligence, knowledge representation, knowledge acquisition, and linguistics [8].

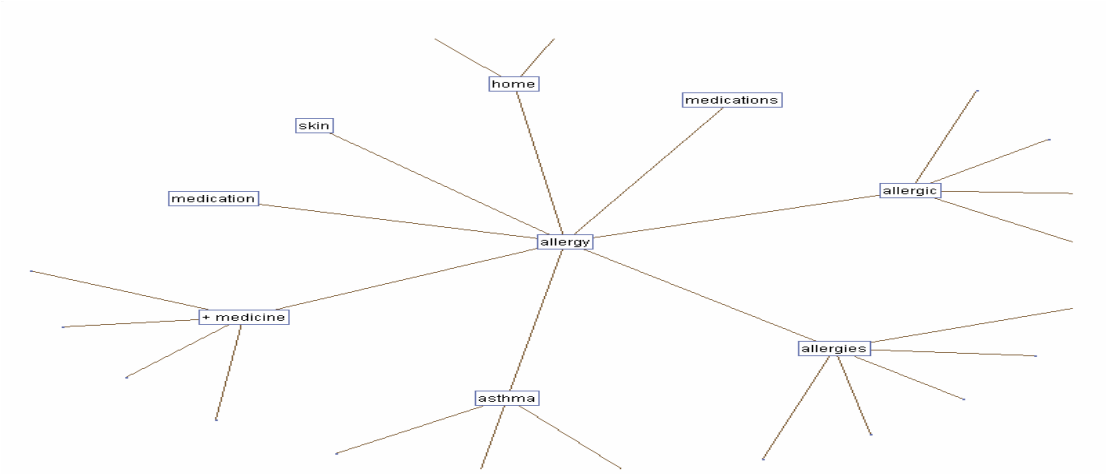
Text Miner provides the user with the concept links feature. Not all terms will have potential links. Note that the first time the user asks for concept links, a column appears in the term window containing the number of potential links. This column can be sorted by clicking on the name “links” so that the user can find the terms with the maximum number of links [7]. The concept link appears in a browser window as an html document. The link can be animated by moving the mouse cursor. All of the lines connected to words are themselves connected to related words that can be discovered by moving the mouse around. Clicking the mouse on one of the words will connect to another browser window providing the documents that comprise the link. The algorithm used to define the concept links is that of association rules [7]. The concept links for “pollen” and “allergy” are given in Figures 1 and 2 respectively.

Figure 1. Concept links for the term “pollen”



Several of the links are related to the pollen, showing different kinds of pollens such as dust and air. Some information is also given on how to avoid pollen. Still one more association is with symptoms (sneezing, eye).

Figure 2. Concept links for the term "allergy"



Note that some of the links are about the type of allergies and allergy medications. One interesting link is a connection between allergy and asthma.

DISCUSSION

The results show that most of the medical information online is about the type of allergy medication drugs, while there are almost no reports of that kind of information on Medline. Today, patients are able to bring to their doctors a large quantity of online information about their illnesses and treatment options. They are prepared to discuss all of it with their physicians. Doctors must accept that this information is readily available to their patients.

Text miner provides a powerful technique for the medical field. It can perform an automatic search of the internet to find thousands of documents and articles that are related to a specific topic. It helps analyzing those documents in order to find patterns and relationships. It is not necessary to read every document. This is a good saving of time and money. Text miner also provides a better use of medical databases such as Medline. It is impossible to investigate and examine manually all the documents related to a particular subject. The keyword search engines have their limitations. Text miner is able to find, classify and cluster documents based on some similarity measures.

REFERENCES

1. www.PupMed.gov.
2. Internet Health Ressources: Health searches and email have become more commonplace, but there is room for improvement in searches and overall Internet access.
3. *Empirical studies assessing the quality of health information for consumers on the world wide web: a systematic review.* . Jama. 2002;287(20):2691-2700.
4. www.hitiweb.mitreek.org/docs/policy.html.
5. *Instructions for authors: Criteria for Manuscripts.*
6. SAS, *SAS course notes.*
7. cerrito, P., *Introduction to use of text miner software.*
8. Gaines, R.K.a.B.R., *Embedded Interactive Concept Maps in Web Documents.* World Conference of The Web Society.

CONTACT INFORMATION

Chakib Battioui
University of Louisville
Louisville, KY
(502) 852-6240
c0batt01@gwise.louisville.edu