Intersection Graphs for Text Analysis

Elizabeth Leeds* David J. Marchette*

August 19, 2004

Abstract

We use a word weighting for text documents that allows for creating individual stopper lists for each document. This method of feature selection is novel in that a word may be kept as an important feature of one document but discarded for another document. Intersection Graphs are used to determine document similarity and to select document specific important words. The characteristics of these graphs are then investigated. Various similarity measures are explored.

Keywords Text processing, intersection graphs, dimensionality reduction, stopper list.

1 Introduction

In document clustering and classification, a document is usually represented as a vector in high dimensional space where each dimension corresponds to a word in the lexicon. Each element of the vector is a weight assigned to the document for the corresponding word. This is referred to as a "bag-of-words" encoding; the order of the words in the document is lost, only a measure of the frequencies is kept. One widely used weighting schemes is TFIDF (term frequency inverse document frequency) of Salton and Buckley [1988] which assigns weights proportional to the number of times the word is in the document and inversely proportional to the fraction of documents that contain the word (weight = term frequency $*$ *log* (inverse document frequency)). Words that are not contained in the document have a weight of zero as do words that are contained in every document in the corpus. All other words have a positive weight. In this way, the documents are represented as vectors and for each class, a prototype vector is constructed as the sum of all documents of that class in the training set. A new document is said to belong to the same class as the prototype vector to which it has the smallest cosine distance.

In this work, we use a weighting method motivated by mutual information similar to Lin and Pantel [2002], Pantel and Lin [2002]. Unlike TFIDF, this

*Naval Surface Warfare Center, Code B10, Dahlgren, VA 22448. {marchettedj, leedsem}@nswc.navy.mil

representation does not map the documents into a vector space (see Section 2). Nevertheless, it gives us a measure of the importance of words within documents that allows for a document specific extraction of representative words. While Lin and Pantel [2002] use pointwise mutual information between a word and a concept (see Manning and Schütze [2002] p.178), we measure mutual information between a word and a document. By using this weighting, we can retain words that are important for specific documents while ignoring the same words in documents for which they do not provide information.

In text analysis, words in documents are usually stemmed, removing prefixes and suffixes. After stemming, words such as 'walked', 'walking', and 'walks' are all represented by the same root word. In this work, we use the Porter stemmer (Porter [1980]).

In addition to stemming, a "stopper" list is often employed. This is a list of words such as 'and', 'of', and 'the', that are assumed not to be of value in the analysis. Many weighting methods assign such words a weight of zero. For example, TFIDF assigns a weight of zero to any word that occurs in every document. However, consider the case of a corpus of documents about circuits. Here the word 'and' may appear with a higher frequency in a subset of articles about AND gates. Perhaps the word should not be eliminated in these articles.

Since the word weighting that we employ is based on the frequency of a word in a document compared to the frequency of the word in an *average* document in the corpus (see Section 2), we can retain words that are significant given the corpus and document. We show how to create a stopper list that is not only corpus dependent but also document specific. That is, each document has its own subset of words that are eliminated. We use intersection graphs to demonstrate a principled method for selecting the threshold resulting in the best representation of the documents.

We perform an experiment on a corpus of documents taken from Science News. For each document, we find the weight of each word based on the mutual information. Then we find a threshold for which we ignore words with lower weight. This reduces the list of words used to represent each document. We use an intersection graph to represent the similarity between documents. Several thresholding schemes and similarity measures are investigated.

2 Notation and Word Weighting

Let c_{dw} be the number of times that the word w has occurred in the document d and let $N = \sum_{ij} c_{ij}$ be the total number of words (counting duplicates) in the corpus $\mathcal{D}$. Let $f_{dw} = c_{dw}/N$. Then the mutual information between document d and word w is given by

$$
\begin{aligned}
m_{dw}^{\mathcal{D}} &= \log\left(\frac{f_{dw}}{\sum_j f_{jw}\sum_i f_{di}}\right) \\
&= \log\left(\frac{c_{dw}}{N\sum_j \frac{c_{jw}}{N}\sum_i \frac{c_{di}}{N}}\right)
\end{aligned}
\tag{1}
$$

The superscript $\mathcal{D}$ is included to make clear that the weight is dependent on the corpus but will be omitted unless needed for clarity. Let $N_d = \sum_i c_{di}$ be the number of words (counting duplicates) in document d. Then

$$
m_{dw} = \log\left(\frac{c_{dw}}{\frac{N_d}{N}\sum_j c_{jw}}\right)
$$

Let $c_{\mathcal{D}w} = \sum_j c_{jw}$ be the number of times that the word w appears in the corpus $\mathcal{D}$. Then

$$
m_{dw} = \log\left(\frac{\frac{c_{dw}}{N_d}}{\frac{c_{\mathcal{D}w}}{N}}\right) \tag{2}
$$

Equation 2 shows that a positive weight will be assigned to a word if the word appears in the document at a higher frequency than it appears in the corpus on average. Note that by definition, any word in the lexicon that does not appear in a document has a weight of $-\infty$. This is unlike other weighting methods that assign nonexistent words a weight of zero. Therefore, we are not representing the documents in a vector space and a cosine distance between documents does not have meaning. However, as shown in Section 3, we represent each document as a list of words, apply a threshold to eliminate words with low information, and investigate different measures of similarity by measuring intersections of the word lists.

3 Intersection Graphs

A graph is a vertex set along with a set of unordered pairs of vertices, called the edge set. Denote the vertex set of the graph G by $V(G)$ and the edge set by $E(G)$. Then an edge exists between vertices v and w if $vw \in E(G)$. The number of vertices is called the *order* of the graph and the number of edges is referred to as the *size* of the graph. For a graph theory reference, see Bollobás [1998].

The graph G is an intersection graph if a set S_v can be assigned to each vertex $v \in V(G)$ so that $vw \in E(G)$ exactly when $S_v \cap S_w \neq \emptyset$. Any graph can be represented by an intersection graph (Karonski et al. [1999]). We use an intersection graph to represent relationships between documents by allowing each document to be a vertex in the graph and placing an edge between vertices when the corresponding documents exceed some similarity measure. Let v_j be

the vertex representing document j and let S_j be the set of words representing the document. Then $v_i v_j \in E(G)$ if S_i and S_j are similar, meaning documents i and j share important words. This is explained in Section 4.1.

3.1 Thresholding

Let S_j be the set of words contained in document j and let $d_j = \{m_1, m_2, ..., m_{|S_j|}\}$ be the ordered set containing the weights for each word in S_j. Consider two types of thresholding. The first method of thresholding eliminates words with weights below a set threshold and retains words with weights above the threshold. The weights are not retained, they are only used to determine which words to keep in the word list representing each document.

$$T(m, \tau) = \begin{cases} 0 & if \quad m < \tau \\ 1 & if \quad m \geq \tau \end{cases} \tag{3}$$

Another thresholding eliminates words which have a weight below a set threshold and retains the weights above the threshold in order to retain importance about the corresponding words in the documents.

$$t(m, \tau) = \begin{cases} 0 & if \quad m < \tau \\ m & if \quad m \geq \tau \end{cases}, \text{ for } \tau \geq 0. \tag{4}$$

Equation 4 can be modified to allow for negative τ, but we will assume $\tau \geq 0$. With the first type of thresholding, only the intersections of the word lists can be used in the determination of edges while in the latter case, a weighted intersection can be used. This is shown in Section 3.2.

3.2 Measures of Similarity

If two documents are determined to have high similarity, their corresponding vertices are connected by an edge. In the simplest case, let $v_i v_j \in E(G)$ if

$$|S_i^\tau \cap S_j^\tau| \geq k \text{ for some } k \in Z^+$$

where S_j^τ be the set of words contained in document j after applying threshold value τ. We can omit the superscript τ unless needed for clarity. This measure of similarity is problematic for documents of differing lengths since document size affects intersection size. To account for this, let $v_i v_j \in E(G)$ if

$$\frac{|S_i \cap S_j|}{\sqrt{|S_i||S_j|}} \geq q \in \mathbb{R}^+. \tag{5}$$

This measure is equivalent to applying the thresholding of Equation 3, representing the documents in the space $\{0,1\}^{|\mathcal{L}|}$ where $|\mathcal{L}|$ is the size of the lexicon, and measuring the cosine similarity.

We can retain more information by using the thresholding of Equation 4, representing the document as a vector $d_j \in \mathbb{R}^{|\mathcal{L}|}$ where elements of the vector that correspond to missing words (or words that have been eliminated by

thresholding) have a value of zero, and measuring cosine similarity. In this case, let $v_i v_j \in E(G)$ if

$$\frac{d_i \cdot d_j}{||d_i||||d_j||} \geq q \in \mathbb{R}^+. \tag{6}$$

4 An Experiment: the Science News Corpus

We performed an experiment using a collection of articles from the Science News from 1994 through 2002. The corpus was collected from the Science News website on December 19, 2002 and consists of 1160 articles that were later manually classified using 8 classes. Of these 1160 documents, 1047 have been assigned to a unique class. The uniquely classified documents are used throughout this paper. The classes and number of articles in each are given in the following table. Even though these articles were classified as belonging to a single class, there is some natural overlap between classes. For example, there are still some documents that should be classified as Behavior *and* Medicine or as Physics *and* Math & Computers. This accounts for some of the edges in the intersection graph that connect documents of different classes.

Classification	# of Docs
Anthropology	54
Astronomy	121
Behavior	72
Earth	137
Life	205
Math & Computers	60
Medicine	280
Physics	118
TOTAL	1047

4.1 Procedure

The corpus $\mathcal{D}$ is the set of 1047 uniquely classified documents. Let S_j be the set of words contained in document j. Let $d_j = \{m_1, m_2, ..., m_{|S_j|}\}$ be the ordered set containing the weights for each word in S_j where $m_i = m_{di}$ as given by Equation 2. Apply the threshold τ so that S_j^τ is the truncated list of words representing document j.

Results are shown using thresholding from Section 3.1 and similarity measures from Section 3.2. We look at how varying the threshold affects the performance. In addition, we look at the performance using a scaled intersection of Equation 5 and the weighted intersection of Equation 6. In either procedure, the number of edges is fixed in order to compare performance[1].

[1]For a fixed graph size of N, the top N similarity values are used to create the edges in the graph. However, the graph size can be larger than N if there are several values equal to the N^{th} highest value.

We use the scaled intersection of Equation 5 to define the edge set of a graph and incrementally adjust the threshold τ in order to determine the best list of words for each document given the corpus. The graph of size N, $G(N,\tau)$, is defined by letting the N largest values of

$$\frac{|S_i^\tau \cap S_j^\tau|}{\sqrt{|S_i^\tau||S_j^\tau|}}$$

specify the edge set of the graph. That is, $v_i v_j \in E(G(N,\tau))$ if documents i and j have one of the N largest similarity values.

By increasing τ, the list of words used to represent the documents decreases. We show that using a subset of words from each document provides better results and we show how to find the value of τ that defines this subset for each document.

4.2 Results

Figure 1 shows edges on a graph of size 2000 using the corpus of 1047 articles. The word lists were thresholded at $\tau = 1.0$ and the scaled similarity measure of Equation 5 was used. As expected, documents in the same class tend to be more connected than documents in different classes. Also as expected, there are many connections between similar topics such as Physics and Astronomy and Behavior and Medicine. Also note that there are often edges from many documents of one class to one document of another class.

Varying the threshold while keeping the graph size constant shows the influence of the threshold on the performance of the intersection graph to connect articles within the same class. Figure 2 shows the fraction of edges between documents of different classes as τ is varied. This is created using all 1047 documents, a graph size of 2000, and the scaled intersection of Equation 5.

Figure 2 shows that increasing τ up to an optimal value has the effect of removing noise, but past this value the thresholding begins to remove information. The optimal value of τ results in the best subsets of the word lists for the documents. This approach to feature selection allows for word lists which are corpus dependent.

Figure 3 shows the performance using the weighted intersection of Equation 6 along with the scaled intersection of Equation 5. The figure shows better performance using the scaled intersection. The weights appear to be too noisy to contribute to the performance of the intersection graph. However, we have observed that the weights provide useful information in other analyses such as the projection of the spectrum of the graph.

5 Adding Documents to the Corpus

Let the document j be in the corpus $\mathcal{D}$. Suppose we add a new set of documents, $\mathcal{D}_1$, to $\mathcal{D}$ and measure the change of m_{jw} under this change in corpus. The

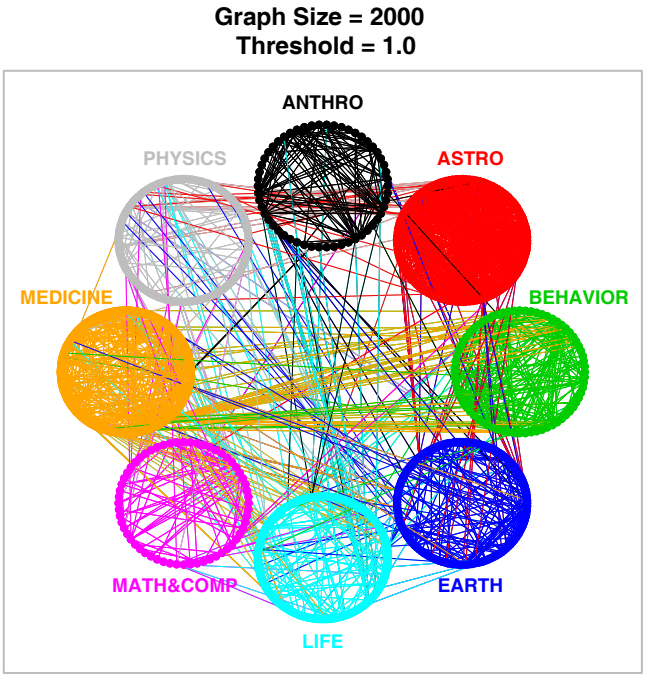

Figure 1: Intersection graphs of the 1047 uniquely classified Science News documents. The lists are thresholded at 1.0 and a scaled similarity measure is used. As can be expected, documents in the same class tend to be more connected than documents in different classes. Also as expected, there are many connections between similar topics such as Physics and Astronomy and Behavior and Medicine. Also note that there are often edges from many documents of one class to one document of another class.

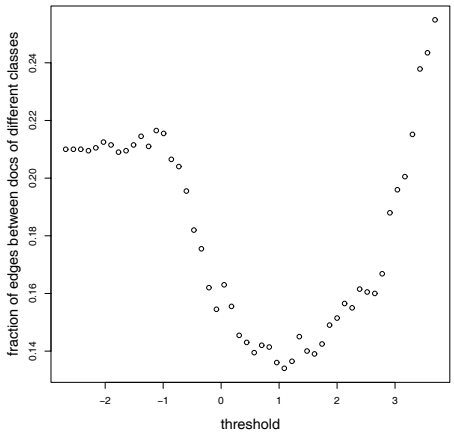

Figure 2: The fraction of edges between documents of different classes as the threshold on the word list, τ, is varied. All 1047 documents were used and the graph size was fixed at 2000. The scaled intersection of Equation 5 was used.

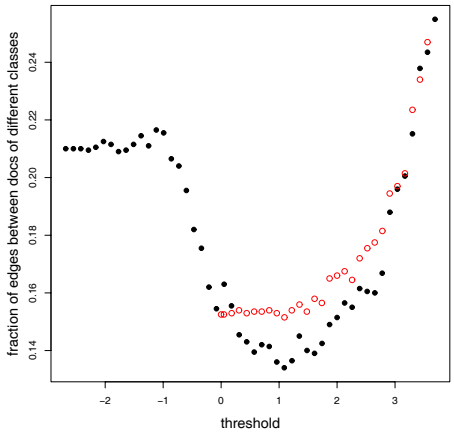

Figure 3: The fraction of edges between documents of different classes as the threshold is varied. All 1047 documents were used and the graph size was fixed at 2000. The solid black points show the performance using the scaled intersection of Equation 5. The open red points show the performance using the weighted intersection of Equation 6.

change in the mutual information of word w in document j under the addition of the set of documents $\mathcal{D}_1$ is

$$\begin{aligned}\Delta_{m_{jw}}^{\mathcal{D}_1} &= m_{jw}^{\mathcal{D}\oplus\mathcal{D}_1} - m_{jw}^{\mathcal{D}} \\ &= \log\left(\frac{c_{jw}}{N_j}\frac{N_{\mathcal{D}\oplus\mathcal{D}_1}}{c_{\mathcal{D}\oplus\mathcal{D}_1 w}}\right) - \log\left(\frac{c_{jw}}{N_j}\frac{N_{\mathcal{D}}}{c_{\mathcal{D}w}}\right) \\ &= \log\left(\frac{c_{\mathcal{D}w}}{N_{\mathcal{D}}}\frac{N_{\mathcal{D}\oplus\mathcal{D}_1}}{c_{\mathcal{D}\oplus\mathcal{D}_1 w}}\right).\end{aligned} \tag{7}$$

This shows that the change in the mutual information for the word w does not depend on the document j. That is, for each word the mutual information will change by the same amount across all documents. All words that appear in the original corpus $\mathcal{D}$ can be computed in the new corpus $\mathcal{D}\oplus\mathcal{D}_1$ by

$$\begin{aligned}m_{jw}^{\mathcal{D}\oplus\mathcal{D}_1} &= m_{jw}^{\mathcal{D}} + \Delta_{m_w}^{\mathcal{D}_1} \\ &= m_{jw}^{\mathcal{D}} + \log\left(\frac{c_{\mathcal{D}w}}{N_{\mathcal{D}}}\frac{N_{\mathcal{D}\oplus\mathcal{D}_1}}{c_{\mathcal{D}\oplus\mathcal{D}_1 w}}\right).\end{aligned} \tag{8}$$

If a word is less common in the new corpus than it was in the old corpus, the weight will increase. Equation 8 provides a simple method for updating the word weights as the corpus grows.

Figure 4 shows the effect of adding documents of new classes to the corpus. The optimal value of τ increases as more documents are added, as does the range of weights. For class dependent words w in documents d of class K, m_{dw} increases as you add documents of different classes. Consequently, the optimal value of τ will increase as well.

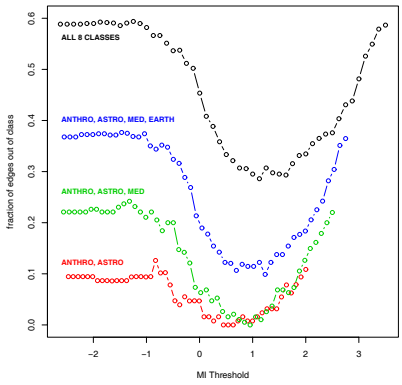

Figure 4: As documents are added to the corpus, the optimal value of τ increases.

6 Summary

We illustrated a method of word weighting that is easily updated with additions to the corpus. Also, we have shown the importance of thresholding the word lists to reduce the noise in the data. This results in a stopper list for each document as opposed to a stopper list for the corpus as a whole. However, choosing the optimal value of the threshold τ is not as easy in practice if the task is clustering and thus the true classes are not known. In future work, attempts will be made to find a relationship between the characteristics of the corpus and the optimal value of τ.

7 Acknowledgments

This work was supported by the Office of Naval Research In-house Laboratory Independent Research (ILIR) funding.

References

Béla Bollobás. *Modern Graph Theory*. Springer-Verlag, New York, 1998.

Michal Karonski, Edward R. Scheinerman, and Karen B. Singer-Cohen. On random intersection graphs: The subgraph problem. *Combinatorics, Probability and Computing*, 8:131–159, 1999.

Dekang Lin and Patrick Pantel. Concept discovery from text. In *Proceedings of the Conference on Computational Linguistics*, pages 577–583, 2002.

C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Massachusetts, 2002.

Patrick Pantel and Dekang Lin. Discovering word senses from text. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 613–619, 2002.

M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.