The Clustering Power of Low Frequency Words in Academic Webs[1]

Liz Price
School of Computing and Information Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1SB, UK. E-mail: Liz.Price@wlv.ac.uk
Tel: +44 1902 321470 Fax: +44 1902 321859
Mike Thelwall
School of Computing and Information Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1SB, UK. E-mail: m.thelwall@wlv.ac.uk
Tel: +44 1902 321470 Fax: +44 1902 321478

The value of low frequency words for subject-based academic web site clustering is assessed. A new technique is introduced to compare the relative clustering power of different vocabularies. The technique is designed for word frequency tests in large document clustering exercises. Results for the Australian and New Zealand academic web spaces indicate that low frequency words are useful for clustering academic web sites along subject lines; removing low frequency words results in sites becoming, on average, less dissimilar to sites from other subjects.

Keywords **web mining, scientific web intelligence**

1. Introduction

An established research problem in information science is that of identifying relationships in the communication of knowledge, both within and between fields and disciplines. Relational bibliometrics (Borgman & Furner, 2002), for example, comprises a set of quantitative techniques for analysing relationships through scholarly documents, typically using citations or document text for data and producing diagrams or other visualizations as outputs (c.f. Börner, Chen, & Boyack, 2003). Depending upon the scale of the analysis, relational studies can help researchers understand their own field structure (White & McCain, 1998), its relationship to other fields (Leydesdorff, 2004), the interaction between different fields (Heimeriks, Hörlesberger, & van den Besselaar, 2003) or the broad structure of science (Glänzel & Schubert, 2001; Small, 1999). Academic web sites are an attractive alternative data source for relational studies because they can be timelier than journals. In extreme cases, web sites announcing research projects may be placed online before the projects start, but publications resulting from the research may appear years after the projects have finished, because of publication delays. The new field of scientific web intelligence (SWI) has the goal of producing effective subject-based visualizations of collections of academic web sites (Thelwall, 2004a). Following previous practice, an 'academic web site' will be operationalized as a collection of URLs sharing a common domain name (Thelwall, 2004b). In some cases these will be multiple web sites or only parts of larger web sites but, on average, this is a useful definition (Bharat, Chang, Henzinger, & Ruhl, 2001; Björneborn, 2001; Thelwall, 2002).

Effective subject-based clustering of web sites is a pre-requisite for effective subject-based visualizations. Web clustering may be based upon web page text (e.g., Kobayashi, & Aono, 2003) or links (e.g. Flake, Lawrence, Giles, & Coetzee, 2002). In this paper we focus on the former, i.e. assuming that web sites covering similar academic subjects will tend to possess similar words and (relative) word frequencies. When clustering any kind of documents using their text, it is common to give special consideration to high frequency words and low frequency words, an idea originating in Luhn's (1958) intuition that the middle-ranking words in a document are most indicative of its content (c.f. van Rijsbergen, 1979, p10; Lancaster, 1977, p295). Salton pursued word frequency information in order to improve information retrieval systems, for example showing in a test collection that the words with the highest average discriminatory power tended to occur in between 1% and 90% of the documents (Salton, Wong & Yang, 1975). Salton's experiments resulted in the well-known vector space model (Salton & McGill, 1983), which is described below.

Previous research has developed a new technique for differentiating between useful and useless high frequency words, namely vocabulary spectral analysis (Thelwall, 2004b). In this paper, however, we consider

low frequency words. Infrequent words have often been omitted in information retrieval systems with the understanding that they are likely to be spelling mistakes or obscure words with low average document discriminatory power (Salton, Wong & Yang, 1975; van Rijsbergen, 1979; c.f., Weeber, Vos, & Baayen, 2000). This would be highly desirable for SWI applications because eliminating low frequency words reduces the dimensionality of the data for clustering. This can significantly speed up clustering because low frequency words are predominant in document collections, in terms of *distinct* word occurrences (Zipf, 1949). A significant proportion of low frequency words in academic webs are not errors, however (Thelwall, 2004c). Hence, it is an open question as to whether they will be useful to help cluster academic web sites. Note that there are other approaches to reducing the dimensionality of the data. Latent Semantic Indexing (LSI) (Deerwester, Dumais, Furnas, et al., 1990) is one logical choice. This operates by identifying underlying patterns in word use, allowing documents to be indexed against the patterns rather than individual words. The complexity and time taken to run an LSI analysis has lead to a similar, but simpler, approach (Kohonen, Kaski, Lagus et al., 2000), which selects words at random from documents for indexing. Neither of these is appropriate for our task, however, because the ultimate objective of SWI domain visualisation is to report *causes* of similarity to users. LSI is also undesirable because of the difficulty, in practice, of mapping LSI patterns (eigenvectors) to meaningful concepts.

In this paper, we assess whether low frequency words are helpful for academic web site clustering, using an experiment on the web sites of the universities in Australia and New Zealand. To help answer the question, we introduce a new technique designed to assess the relative clustering power of different sets of words (sub-vocabularies).

2. A clustering power assessment technique

2.1 Initial partial clustering

The two clustering power estimation techniques used in this paper (sections 2.2 and 2.4) require a set of partially correct subject-based clusters to serve as benchmarks against which to assess the subject-clustering power of different vocabularies. These are subsets $C_1, C_2, ...C_p$ of the whole document space X. For the technique to work, all that is required is that documents inside each C_i are more likely to be about the same subject as other documents in C_i than the same subject as documents outside C_i. The subsets $C_1, C_2, ...C_p$ do not need to be disjoint (non-overlapping) or to include every document in X. In general, however, larger and more accurate clusters make the technique more powerful, each C_i ideally containing all documents relevant to a single subject.

The objective of most clustering applications is only to identify clusters in the data. For SWI, in contrast, the desired *types* of clusters are known in advance: they should be subject-based. Hence, an acceptable method for identifying $C_1, C_2, ...C_p$ would be to use human classifiers to classify a subset of (domain name-based) web sites from the data set. In fact it is possible to semi-automate this process by exploiting regularities in domain names. For example, domain names containing 'math' tend to be owned by mathematicians. A human classifier can therefore build up a set of rules (regular expressions) with which to check all domain names in the data set (e.g. 'if the domain name contains "math" then assign the domain to the mathematics cluster'), classifying only those that match one of the expressions. This is an iterative process, with the human classifier checking the results of each automatic classification attempt and modifying the bank of rules accordingly. The set of expressions is nation-specific but can be ported to different countries with minimal changes.

The result of the process described above is a list of domains that have been allocated to specific subjects through matching one of the rules. In any application, the majority of domains will probably not match any of the rules and therefore will not be allocated to a subject, being left unclassified. As described below, the partial nature of the clustering is not a problem.

2.2 Cluster Matching Power

In this section, a technique is introduced to assess how well a set of subject clusters matches a set of automatically calculated clusters. Let V be the vocabulary for the document space X, so that V is the set of all distinct words that have been extracted from the documents in X. Let $S \subset V$ be a sub-vocabulary. Clustering the

document space X based upon S means performing the clustering operation after discarding all words that are not in S. One way to use the partial clusters $C_1, C_2, ...C_p$ to assess the power of S to cluster a set of documents is to automatically cluster the full set of documents X to generate a new set of clusters $C'_1, C'_2, ...C'_p$ and then to use a measure to assess how far the manually identified subject clusters match the automatically generated clusters. The measure that we shall use to assess this match is the average number of automatic clusters C'_k that each manually identified subject cluster C_k spreads across. This will be called the cluster matching power (CMP) of S, denoted *CMP(S)*. Lower CMP values indicate a more powerful subvocabulary because the clusters it generates are less spread out amongst the subject clusters.

A disadvantage of CMP values is that they will not be highly sensitive to small changes in clustering power. An alternative would be to use a finer-grained method such as probabilistic clustering or soft *k*-means (Baldi, Frasconi, & Smyth, 2003; Hand, Mannila & Smyth, 2001). A similar motivation has lead to new measures for information retrieval effectiveness (Della Mea & Mizzaro, 2004). The more sensitive clustering methods are not appropriate for large document clustering exercises, however, unless sufficient time and computing power are available. We introduce below a new vector space model (VSM) metric designed to identify small differences in clustering tendency that might not be reflected in actual clusters found. Before introducing the new metric, the vector space model is described.

2.3 The Vector Space Model

The vector space model is a standard information retrieval approach for document representation and for use in document relevance ranking (Salton & McGill, 1983; Baeza-Yates & Ribeiro-Neto, 1999). The model interprets all documents as "bags of words" in no particular order. Each document is represented only by a list of the words that it contains and a list of how often each word occurs (which are later converted to weights using a mathematical formula). Clearly, much information is lost because the order of the words is not recorded. For example, if the words "New" and "Mexico" occur in a document then it is much more likely that the document relates to New Mexico if the words are known to be consecutive. Nevertheless, the vector space model is a useful representation of documents because it allows efficient searching and clustering.

When applying the VSM to a set of documents, the first step is to construct a vocabulary, a list of all words found in any of the documents. The individual documents are then converted to word frequency vectors (simple lists of numbers) by recording the frequency of each of the words in the vocabulary. Normally, any document will only contain a small minority of the words in the vocabulary and therefore most of the word frequencies will be zeros. The VSM exploits word frequency information in order to generate weights for all of the words in a document. These estimate the relevance of each word to the document. A mathematical formulation follows.

Let X be a set of *n* documents. Let *m* denote the total number of unique words, and let n_i be the number of documents containing word *i*. Let f_{ij} be the frequency of word *i* in document *j* with $f_{j\max}$ being the maximum frequency of any word in document *j*.

The standard VSM weighting for a word *i* in document *j* is $w_{ij} = \frac{f_{ij}}{f_{j\max}} \log\left(\frac{n}{n_i}\right)$. Thus, words in a document are weighted highly if they occur in few documents (i.e., n_i is low, so $\frac{n}{n_i}$ is high and hence $\log\left(\frac{n}{n_i}\right)$ is high) and have a high frequency relative to other words in the document (i.e. $\frac{f_{ij}}{f_{j\max}}$ is high). The VSM represents documents by word frequency weight vectors, so that document *j* is represented by the vector $(w_{ij})_{i=1...n}$. The distance between two documents is commonly measured with the cosine measure, defined below for documents *j* and *j'* (Baeza-Yates & Ribeiro-Neto, 1999).

$$\frac{\sum_{i=1}^{n} w_{ij} w_{ij'}}{\sqrt{\sum_{i=1}^{n} w_{ij}^2} \sqrt{\sum_{i=1}^{n} w_{ij'}^2}} \qquad (1)$$

The cosine measure gives values between 0 and 1, with documents that contain similar words tending to have a similarity close to 1, and documents with few words in common tending to have a similarity close to 0. We calculate the (cosine) similarity $d_S(j, j')$ of pairs of documents *j* and *j'* over sub-vocabularies *S*, by using only words in *S* for all calculations, as shown below.

$$d_S(j, j') = \frac{\sum_{i \in S} w_{ij} w_{ij'}}{\sqrt{\sum_{i \in S} w_{ij}^2} \sqrt{\sum_{i \in S} w_{ij'}^2}} \qquad (2)$$

2.4 VSM-Derived Measures of Clustering Tendency

Salton, Yang and Yu (1975) have developed a technique to assess how far individual words contribute to the clustering of a set of documents. They found that words which increased the dissimilarity of documents within a collection were more likely to be useful for information retrieval purposes, and would also help to cluster similar documents. They introduced a 'term discrimination value', which assesses the extent to which a word makes documents in a collection closer or more distant. This was later criticized by Willett (1985) for being very dependant upon the similarity metric used. In addition, van Rijsbergen (1979, p15) also criticizes the model for attempting to increase the separation of all documents, rather than just the relevant from the non-relevant documents (from an information retrieval perspective). The method described below is able to improve upon the term discrimination value approach through advance knowledge of the desired clusters and so can be used to increase inter-cluster separation relative to intra-cluster separation. Willett's criticism does not apply because the method can be specific to the metric used because the ultimate aim is clustering.

For a given sub-vocabulary S and cluster C, the power of S to discriminate between members of C and non-members of C is the average similarity of distinct pairs of documents in C (using equation 2) minus the average similarity of documents in C with documents outside of C, as shown below. *CT(S,C)* is the clustering tendency of the cluster C using the subvocabulary S.

$$CT(S, C) = \frac{\sum_{j \in C} \sum_{j' \in C \setminus \{j\}} d_S(j, j')}{|C|(|C| - 1)} - \frac{\sum_{j \in C} \sum_{j' \in \overline{C}} d_S(j, j')}{|C||\overline{C}|} \qquad (3)$$

A normalized version of this calculation is needed because the average differentiating power between documents may not be uniform across sub-vocabularies. For example, a statistical by-product of decreasing the vocabulary is to make documents appear smaller and this tends to make similarity measures larger. *NCT(S,C)* is the normalized version of the above formula. Normalisation is achieved by dividing by the average similarity of documents within a cluster to all other documents.

$$NCT(S, C) = \frac{CT(S, C)}{\frac{\sum_{j \in C} \sum_{j' \in C \setminus \{j\}} d_S(j, j') + \sum_{j \in C} \sum_{j' \in \overline{C}} d_S(j, j')}{|C|(|C| - 1) + |C||\overline{C}|}} \qquad (4)$$

For a set of user clusters, the average of these differences *ANCT(S)* can be used as a measure of the clustering effectiveness of the sub-vocabulary S, for the pre-selected clusters $C_1, C_2, ...C_p$

$$ANCT(S) = \frac{1}{p} \sum_{k=1}^{p} NCT(S, C_k) \qquad (5)$$

3. Experimental Design

We assess the clustering power of low frequency words in academic webs using two data sets: New Zealand university web site, and Australian university web sites. We test clustering power with a range of different word

frequency thresholds and using two different methods. First, the clustering is assessed using CMP values to test whether removing low frequency words makes the clusters worse, i.e. whether the automatically generated clusters become more spread out amongst the human-identified subject clusters. Second, the more sensitive ANCT is used to assess whether the clustering tendency decreases when low frequency words are removed, i.e. whether documents in the human-identified subject clusters become less dissimilar to documents outside of their subject cluster.

4. Method

The web spaces of all the universities of New Zealand were crawled in December 2003 and those of Australia in February 2004. A specialist information science web crawler was used, designed to eliminate duplicate pages and ignore mirror sites via a manually maintained banned list (Thelwall, 2001; 2003). Basic word stemming of plurals was used, but, because of the SWI task, extended stemming (e.g. Porter's Algorithm (Porter, 1980)) was not used. As mentioned above, academic web sites are equated with unique domain names. The sites in the study are all subdomains of the university domain names in each country, provided that (a) they have been found by the crawler and (b) they contain at least one non-stopword. All pages with a common domain name were compiled into a single vector for VSM purposes. University home sites were excluded (e.g. www.acu.edu.au) because these are typically of a general nature and not subject-specific. Library, student support, administrative, email, general information, and site-wide courseware sites were also excluded. vocabulary spectral analysis (VSA) (Thelwall, 2004b) was subsequently used to compile a list of high frequency stop words to exclude. VSA is a computer-assisted technique for the human identification of high frequency words that cluster collections of documents in undesired ways (e.g. by university instead of by subject). For Australia, there were 3,780 domains containing a total of 346,417,602 words (1,364,734 unique words), and for New Zealand there were 663 domains containing a total of 41,300,201 words (337,225 unique words).

Subject-based clusters (of academic web sites) were identified using the initial partial clustering method described above (Section 2.1). For Australia this gave 29 subject-based clusters of at least 2 domains, containing a total of 709 clustered domains out of 3,780. For New Zealand this gave 24 subject-based clusters of at least 2 domains, containing a total of 218 clustered domains out of 663. In both cases, the remaining unclustered domains were retained in a single large general cluster. These clusters are the pre-selected clusters $C_1, C_2, ...C_p$ used in equation 5 and in Section 2.2 for Cluster Matching Power calculations.

The sites were clustered using the k-means algorithm with random starting points and one more cluster than the number of clusters automatically identified. This produces the clusters $C'_1, C'_2, ...C'_p$ needed for the Cluster Matching Power calculations (Section 2.2). The clustering was conducted after removing low frequency words at increasing powers of 2, i.e., 2, 4, 8, 16,…. For example, at the threshold of 2, words were rejected that were in less than 2 different domains. The calculation of the vector-space model similarities between pairs of domains took 37 hours for the longest case: Australia with a threshold of 2. Similarities were calculated only once and then the 3,780 x 3,780 or 663 x 663 similarity matrices were fed into the k-means algorithm as raw data, 100 times for Australia and 1000 times for New Zealand in order to obtain reliable averages across different random starting points.

5. Results

Figures 1 and 2 show that eliminating low frequency words reduces the clustering power of the vocabulary for both Australia and New Zealand. In both graphs removing low frequency words reduces ANCT values at all frequency levels. At the lowest level, even increasing the threshold from 2 to 4 lowered the ANCT value, implying that words which occurred in only two or three domains helped to increase clustering tendency. This means that removing the lowest frequency words from the vocabulary tended to make the domains in each cluster less dissimilar to the domains outside of the cluster, as measured using the cosine metric. Figures 3 and 4 show that the quality of clustering for all the different sub-vocabularies stays approximately the same, so that the reduced clustering power of the smaller vocabularies is not significant enough to be transmitted into less effective clustering. The apparent contradiction between the large drops in clustering power evident in figures 1 and 2, and the relatively constant number of automatically generated clusters that the manually identified clusters spread across can be explained by the fact that the clustering power is not high even for the full vocabulary. This can be seen from the relatively high average dispersal of the human identified clusters amongst

the automatically generated clusters. A perfect clustering would give a CMP value of 1: each human classified cluster sitting inside one (possibly larger) automatically identified cluster. Intuitively, as the threshold is raised and the vocabulary decreases, the domains in each human-identified cluster move further apart. Nevertheless, even when the full vocabulary is reached, the domains are the not close enough to significantly clump together, so the decreased distance between domains does not give good automatic clustering or even significantly better clustering. This vindicates the use of the more sensitive ANCT statistic but shows that effective automatic subject clustering of academic webs still has a long way to go. Figure 3 shows the average value over 1000 applications of the *k*-means algorithm, and Figure 4 illustrates the average value over 100 applications.

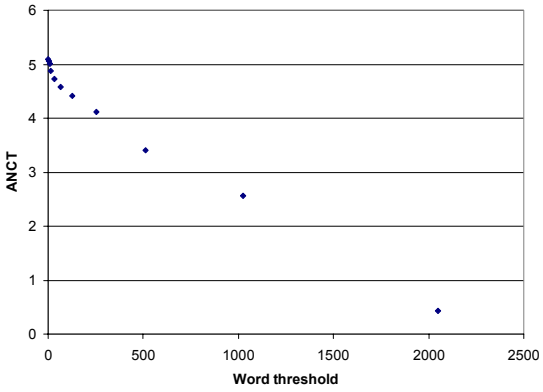

Fig. 1: ANCT values for Australia

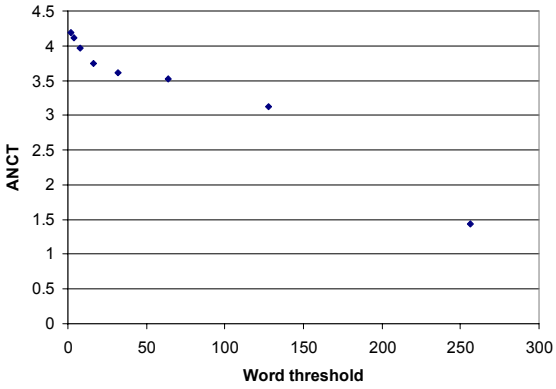

Fig. 2: ANCT values for New Zealand

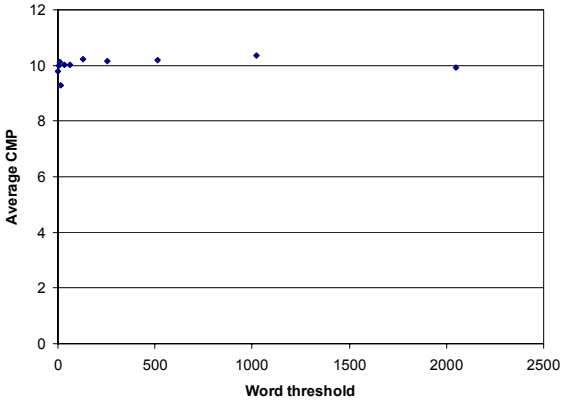

Fig. 3: CMP values for Australia

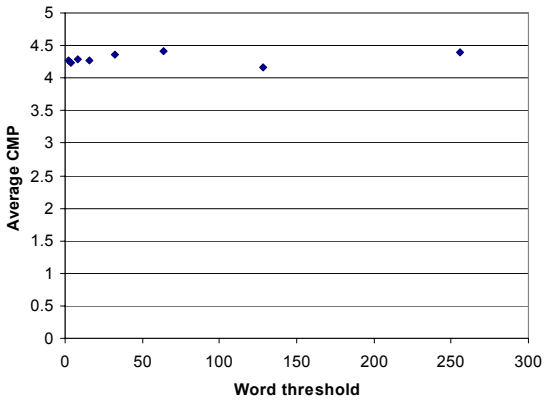

Fig. 4: CMP values for New Zealand

6. Conclusions

We have shown that low frequency words are useful for academic domain clustering. This suggests that a significant proportion of low frequency words contain subject-related information. As a result, it will be undesirable for future SWI research to have a policy of removing all words below any given frequency threshold. Given the computational burden imposed by retaining low frequency words and the fact that automatic clustering gives poor results (figures 3-4), a logical future research direction is the development of artificial intelligence or natural language processing techniques to identify and remove as many words as possible that are of undesired types (e.g. Comeau & Wilbur, 2004).

7. References

Baeza-Yates, R. & Ribeiro-Neto, B. (1999). Modern Information Retrieval. Wokingham, UK: Addison-Wesley.

Baldi, P., Frasconi, P., & Smyth, P. (2003). Modelling the Internet and the Web. Chichester, UK: Wiley.

Bharat, K. Chang, B. Henzinger, M. & Ruhl, M. (2001). Who links to whom: Mining linkage between web sites. In: Proceedings of ICDM 2001, pp. 51-58.

Björneborn, L. (2001). Necessary data filtering and editing in webometric link structure analysis. Royal School of Library and Information Science.

Börner, K. Chen, C. & Boyack, K. (2003). Visualizing knowledge domains. Annual Review of Information Science & Technology, 37, 179-255.

Borgman, C. & Furner, J. (2002). Scholarly communication and bibliometrics. Annual Review of Information Science and Technology 36, 3-72.

Comeau, D.C. & Wilbur, W.J. (2004). Non-word identification or spell checking without a dictionary. Journal of the American Society for Information Science and Technology, 55(2) 169-177.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer T.K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391-407.

Della Mea, V. & Mizzaro, S. (2004). Measuring retrieval effectiveness: A new proposal and a first experimental validation. Journal of the American Society for Information Science, 55(6), 530-543.

Flake, G., Lawrence, S., Giles, C., & Coetzee, F. (2002). Self-organization and identification of Web communities. IEEE Computer, 35, 66-71.

Glänzel, W. & Schubert, A. (2001). Double effort = double impact? A critical view at international co-authorship in chemistry, Scientometrics, 50(2), 199-214.

Hand, P., Mannila, D., & Smyth, H. (2001). Principles of data mining. Boston: MA: MIT Press.

Heimeriks, G. Hörlesberger, M. & van den Besselaar, P. (2003). Mapping communication and collaboration in heterogeneous research networks. Scientometrics, 58(2), 391-413.

Kobayashi, M. & Aono, M. (2003). Vector space models for search and cluster mining. In: Berry, M. Survey of text mining. Springer, New York, pp. 103-122.

Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, J., Paatero, V., & Saarela. A. (2000). Self organization of a massive document collection. IEEE Transactions on Neural Networks, 11(3), 574-585.

Lancaster, E.W. (1977). *Information retrieval systems: Characteristics, testing and evaluation*. New York: John Wiley & Sons.

Leydesdorff, L. (2004). Top-down decomposition of the Journal Citation Report of the Social Science Citation Index: Graph- and factor-analytical approaches. Scientometrics, 60(2), 159-180.

Luhn, H.P. (1958). The automatic creation of literature abstracts, IBM Journal of Research and Development, 2, 159-165.

Porter, M. (1980). An algorithm for suffix stripping. Program, 14 (3), 130-137.

Salton, G. & McGill, J. (1983). *An introduction to modern information retrieval*. New York: McGraw-Hill.

Salton, G., Yang, C.S. & Yu, C.T. (1975). A theory of term importance in automatic text analysis. *Journal of the American Society for Information Science*, 26(1), 33-44.

Salton, G., Wong, A. & Yang, S.S. (1975). A vector space model for automatic indexing, *Communications of the ACM*, 18(11), 613-620.

Small, H. (1999). Visualising science through citation mapping. Journal of the American Society for Information Science, 50(9), 799-812.

Thelwall, M. (2001). A web crawler design for data mining. Journal of Information Science 27(5), 319-325.

Thelwall, M. (2002). Conceptualizing documentation on the Web: an evaluation of different heuristic-based models for counting links between university web sites. Journal of the American Society for Information Science and Technology, 53(12), 995-1005.

Thelwall, M. (2003). A free database of university Web links: Data collection issues. Cybermetrics 6(1). Available: http://www.cindoc.csic.es/cybermetrics/articles/v6i1p2.html.

Thelwall, M. (2004a, to appear). Scientific Web Intelligence: Finding relationships in university webs. Communications of the ACM.

Thelwall, M. (2004b). Vocabulary Spectral Analysis as an exploratory tool for Scientific Web Intelligence. Proceedings of the 8th International Conference on Information Visualisation. Los Alamitos, CA: IEEE, pp. 501-506.

Thelwall, M. (2004c, to appear). Text characteristics of English language university web sites. Journal of the American Society for Information Science and Technology.

van Rijsbergen, C.J. (1979). Information Retrieval. London: Butterworths.

Weeber, M., Vos, R., & Baayen, R.H. (2000). Extracting the lowest-frequency words: pitfalls and possibilities. Computational Linguistics, 26(3), 301-17.

White, H.D. & McCain, K.W. (1998). Visualizing a discipline: an author co-citation analysis of information science, 1972-1995. Journal of the American Society for Information Science, 49(4), 327-355.

Willett, P. (1985). An algorithm for the calculation of exact term discrimination values. Information Processing & Management, 21(3), 225-232.

Zipf, G.K. (1949). Human behavior and the principle of least effort: an introduction to human ecology. Cambridge, MA: Addison-Wesley.