Exploiting Sophisticated Representations for Document Retrieval

Steven Finch
Language Technology Group, HCRC
University of Edinburgh
S.Finch@ed.ac.uk

Abstract

The use of NLP techniques for document classification has not produced significant improvements in performance within the standard term weighting statistical assignment paradigm (Fagan 1987; Lewis, 1992bc; Buckley, 1993). This perplexing fact needs both an explanation and a solution if the power of recently developed NLP techniques are to be successfully applied in IR. A novel method for adding linguistic annotation to corpora is presented which involves using a statistical POS tagger in conjunction with unsupervised structure finding methods to derive notions of "noun group", "verb group", and so on which is inherently extensible to more sophisticated annotation, and does not require a pre-tagged corpus to fit. One of the distinguishing features of a more linguistically sophisticated representation of documents over a word set based representation of them is that linguistically sophisticated units are more frequently individually good predictors of document descriptors (keywords) than single words are. This leads us to consider the assignment of descriptors from individual phrases rather than from the weighted sum of a word set representation. We investigate how sets of individually high-precision rules can result in a low precision when used together, and develop some theory about these probably-correct rules. We then proceed to repeat results which show that standard statistical models are not particularly suitable for exploiting linguistically sophisticated representations, and show that a statistically fitted rule-based model provides significantly improved performance for sophisticated representations. It therefore shows that statistical systems can exploit sophisticated representations of documents, and lends some support to the use of more linguistically sophisticated representations for document classification. This paper reports on work done for the LRE project SISTA, which is creating a PC based tool to be used in the technical abstracting industry.

1 Models and Representations

First, I discuss the general paradigm for document classification, along with the conventions for notation used throughout this document. We have a set of documents $\{x_i\}$, and set of *descriptors*, $\{d_i\}$. Each document is represented in one or more ways in some domain, usually as a set. The elements of this set will be called *diagnostic units* or *predicates*, $\{w_i\}$ or $\{\phi_i\}$. These diagnostic units might be the words comprising the document, or more linguistically sophisticated annotations of parts of the document. They may, in general, be predicates over documents. The representation of the document by *diagnostic units* will be called the *DU-representation* of the document, and for a document x, will be denoted $\mathcal{R}(x)$. From the DU representation of the documents, one or more descriptors are assigned to each of them by some automatic system. This paradigm of description is applicable to much of the work on text classification (and other fields in information retrieval).

This paper assesses the utility of using linguistically sophisticated diagnostic units together with a slightly non-standard statistical assignment model in order to assign descriptors to a document.

2 The Corpus

This paper reports work undertaken for the LRE project SISTA (Semi-automatic Indexing System for Technical Abstracts). This section briefly describes one of the corpora used by this project.

The RAPRA corpus comprises some 212,000 technical abstracts pertaining to research and commercial exploitation in the rubber and plastics industry. To each abstract, an average of 15 descriptors selected from a thesaurus of some 10,000 descriptors is assigned to each article. The frequency of assignment of descriptors varies roughly in the same way as the frequency of word use varies (the frequencies of descriptor tokens (very) approximately satisfies the Zipf-Mandelbrot law). Descriptors are assigned by expert indexers from the entire article and expert domain knowledge, not just from the abstract, so it is unlikely that any automatic system which analyses only the abstracts can assign all the descriptors which are manually assigned to the abstract.

We show a fairly typical example below. It is clear that many of these descriptors must have been assigned from the main text of the article, and not from the abstract alone. Moreover, this is common practice in the technical abstract indexing industry, so it seems unlikely that the situation will be better for other corpora. Nevertheless, we can hope to follow a strategy of assigning descriptors when there is enough information to do so.

Macromolecular Deformation Model to Estimate Viscoelastic Flow Effects in Polymer Melts

The elastic deformation of polymer macromolecules in a shear field is used as the basis for quantitative predictions of viscoelastic flow effects in a polymer melt. Non-Newtonian viscosity, capillary end correction factor, maximum die swell, and die swell profile of a polymer melt are predicted by the model. All these effects can be reduced to generic master curves, which are independent of polymer type. Macromolecular deformation also influences the brittle failure strength of a processed polymer glass. The model gives simple and accurate estimates of practically important processing effects, and uses fitting parameters with the clear physical identity of viscoelastic constants, which follow well established trends with respect to changes in polymer composition or processing conditions. 12 refs.

Original assignment: BRITTLE FAILURE; COMPANY; DATA; DIE SWELL; ELASTIC DEFORMATION; EQUATION; GRAPH; MACROMOLECULE; MELT FLOW; MODEL; NON-NEWTONIAN; PLASTIC; POLYMERIC GLASS; PROCESSING; RHEOLOGICAL PROPERTIES; RHEOLOGY; TECHNICAL; THEORY; THERMOPLASTIC; VISCOELASTIC PROPERTIES; VISCOELASTICITY; VISCOSITY

---

3 Models

Two classes of models for assessing descriptor appropriateness were used. One class comprises variants of Salton's *term-weighting* models, and one is more allied to fuzzy or default logic in so much as it assigns descriptors due to the presence of certain diagnostic units. What is interesting for us is that term weighting models do not seem able to easily exploit the additional information provided by a more sophisticated representation of a document, while an alternative statistical single term model can.

3.1 Term weighting models

The standard term weighting model is defined by chosing a set of parameters $\{\alpha_{ij}\}$ (one for each word-descriptor pair) and $\{\beta_i\}$ (one for each descriptor) so that a likelihood or appropriateness function, $\mathcal{L}$, can be defined by

$$\mathcal{L}(d|\mathbf{w}) = \sum_{w \in \mathbf{W}} \alpha_{wd} + \beta_d \qquad (1)$$

This has been widely used, and is provably equivalent to a large class of probabilistic models (e.g. Van Risjbergen, 1979) which make various assumptions about the independence between descriptors and diagnostic units (Fuhr & Buckley, 1993). Various strategies for estimating the parameters for this model have been proposed (e.g. Salton & Yang, 1973, Buckley 1993, Fuhr & Buckley, 1993). Some of these concentrate on the need for re-estimating weights according to relevance feedback information, while some make use of various functions of term frequency, document frequency, maximum within-document frequency, and various other measurements of corpora. Nevertheless, the problem of estimating the huge number of parameters needed for such a model is statistically problematic, and as Buckley (1993) points out, the choice of weights has a large influence on the effectiveness of any model for classification or for retrieval.

There are so many variations on the theme of term weighting models that it is impossible to try them all in one experiment, so this paper uses a variation of a model used by Lewis (1992c) in which he reports the results of some experiments using phrases in a term weighting model (which has a probabilistic interpretation). Several term weighting models have been tried, but they all evaluate within 5 points of each other on both precision and recall (when suitably tweaked).

The model eventually chosen for the tests reported here was a smoothed logistic model which gave the best results of all the probabilistically inspired term weighting models considered.

3.2 Single term model

In contrast to making assumptions of independence about the relationship between diagnostic units and words, the next model utilises only those diagnostic units which strongly predict descriptors (i.e. have frequently been associated with descriptors) without making assumptions about the independence of diagnostic units given descriptors.

We shall investigate this class of models using probability theory. The main problem with using probability theory for problems in document classification is that while it might be relatively easy to estimate probabilities such as $P(d|w)$ for some diagnostic unit w and some descriptor d, it is not possible to infer much about $P(d|w\Psi)$, where Ψ is some additional information (e.g. the other DUs which represent the document), since these probabilities have not been estimated, and would take a far larger corpus to reliably estimate in any case. The situation gets exponentially worse as the information we have about the document increases. The exception to this rule is when $P(d|w)$ is close to 1, in which case it is very unlikely that additional information changes its value much. This fact is further investigated now.

The strategy explored here is to concentrate on finding "sure-fire" indicators of descriptors, in a somewhat similar manner to how Carnegie's TCS works, by exploiting the fact that with a pre-classified training corpus we can identify sure-fire indicators empirically and "trawl" in a large set of informative diagnostic units for those which identify descriptors with high precision. The basis of the model is the following:

We consider a likelihood function, $\mathcal{L}$ defined by:

$$\mathcal{L}(d|w) = \frac{N_{dw}}{N_w}$$

That is, the number of articles in the training corpus that d was observed to occur with w divided by the number of articles in which w occurred in the training corpus. This is an empirical estimate of the conditional probability, $P(d|w)$. We shall assume (for simplicity's sake) that we have a large enough corpus do reliably estimate these probabilities.

The strategy for descriptor assignment we are investigating is to assign a descriptor d if and only if one of a set of predicates over representations of documents is true. We define the rule $\phi(x) \longrightarrow d$ to be *Probably Correct do degree* ε if and only if $P(d|\phi) > 1-\varepsilon$. We wish to keep the precision resulting from using this strategy high while increasing the number of rules to improve recall. The predicates ϕ we shall consider for this paper will be very simple (they will typically be true iff $w \in \mathcal{R}(x)$ for some diagnostic unit w), but in principle, they could be arbitrarily complex (as they are in Carnegie's TCS). The primary question of concern is whether the ensemble of rules $\{\phi_i \longrightarrow d\}$ retains precision or not. Unfortunately, the answer to this question is that this is not necessarily the case unless we put some constraints on the predicates.

Proposition 1 *Let Φ be a set of predicates with the property that for some fixed descriptor d, $\phi \in \Phi \longrightarrow P(d|\phi) > 1-\varepsilon$. That is each of the rules $\phi_i \longrightarrow d$ is* probably correct *to degree ε.*

The expected precision of the rule $(\bigvee \phi_i) \longrightarrow d$ is at least $1 - \frac{n\varepsilon}{1+(n-1)\varepsilon}$ where n is the cardinality, $|\Phi|$.

Proof:

[Straight-forward and omitted]

This proposition asserts that one cannot be guaranteed to be able to keep adding diagnostic units to improve recall without hurting precision, unless the quality of those diagnostic units is also improved (i.e. ε is decreased in proportion to the number of DUs which are considered). This is unfortunate, but nevertheless the question of how much adding diagnostic units to help recall will hurt precision is an entirely empirical matter dependent on the true nature of P; this proposition is a worst case, and gives us reason to be careful. Performance will be expected to be poorest if there are many rules which correspond to the same true positives, but different sets of false positives. If the predicates are disjoint, for example, then the precision of a disjunction is at least as great as the precision of applying any single rule.

So if we design our predicates so that they are disjoint, then we retain precision while increasing recall. In practice, this is infeasible, but it is feasible to look more carefully at frequently co-occurring predicates, since these will be most likely to reduce precision.[1] The main moral we can draw from the above two propositions is that we must be careful about the case where diagnostic units are highly correlated.

One situation which is relatively frequent as the sophistication of representation increases is that some diagnostic units always co-occur with others. For example, if the document were represented by sequences of words, then the sequence "olefin polymerisation" always occurs whenever the sequence "high temperature olefin polymerisation" occurs. In this case, it might be thought to pay to look only at the most specific diagnostic units since we have if $w_1 \longrightarrow w_2$, then $P(X|w_1w_2C) = P(X|w_1C)$ for any distribution P whatsoever (here, C represents any other contextual information we have, for example the other diagnostic units representing the document). However, if w_1 is significantly less frequent than w_2 estimation errors of $P(d|w_1)$ will be larger for $P(d|w_2)$ for any descriptor d, so there may not be a significant advantage. However, it does give us a

[1] One classic example is the case of the "New Hampshire Yankee Power Plant". In a collection of New York Times articles studied by Jacobs & Rau (1990), the word "Yankee" was found to predict NUCLEAR POWER because of the frequent occurrence of articles about this plant. However, "Yankee" on its own without the other words in this phrase is a good predictor of articles about the New York Yankees, a baseball team. If highly mutually informative words are combined into conjunctive predicates (e.g. "Yankee" $\in x$ & "Plant" $\in x$), and a document is represented by its most specific predicates only, then when "Yankee" appears alone, it will be a good predictor of the descriptor SPORT. This example can also show that the bound described above is tight. Imagine (suspending belief) that each of the five words in the phrase have the same number of occurrences, i, in the document collection without NUCLEAR POWER where they never occur together pairwise, and always occur all together in j true positives of the descriptor. Then the precision of assigning NUCLEAR POWER if any one of them appears in a document is $\frac{j}{j+5i}$, and since ε in this case is $\frac{i}{i+j}$, the bound follows (for the case $n = 5$) with a little algebra.

theoretical reason to believe that representing a document by its set of most specific predicates is worth investigating, and this shall be investigated below.

If one considers a calculus similar to the one described here, but allows ε to limit to 0, then a weak default logic ensues which has been studied by Adams (1975), and further investigated by Pearl (1988).

4 Adding linguistic description

The simplest way of representing a document is as a set or multi set of words. Many people (eg. Lewis 1992bc; Jacobs & Rau 1990) have suggested that a more linguistically sophisticated representation of a document might be more effective for the purposes of statistical keyword assignment. Unfortunately, attempts to do this have not been found to reliably improve performance as measured by recall and precision for the task of document classification. I shall present evidence that a more sophisticated representation makes better predictions from the Single Term model defined above than it does from standard term weighting models.

4.1 Linguistic description

The simplest form of linguistic description of the content of a machine-readable document is in the form of a sequence (or a set) of words. More sophisticated linguistic information comes in several forms, all of which may need to be represented if performance in an automatic categorisation experiment is to be improved. Typical examples of linguistically sophisticated annotation include tagging words with their syntactic category (although this has not been found to be effective for IR), lemma of the word (e.g. "corpus" for "corpora"), phrasal information (e.g. identifying noun groups and phrases (Lewis 1992c, Church 1988)), and subject-predicate identification (e.g. Hindle 1990). For the RAPRA corpus, we currently identify noun groups and adjective groups.

This is achieved in a manner similar to Church's (1988) PARTS algorithm used by Lewis (1992bc), in the sense that its main properties are robustness and corpus sensitivity. All that is important for this paper is that the technique identifies various groupings of words (for example, noun-groups, adjective groups, and so on) with a high level of accuracy. Major parts of the technique are described in detail in Finch, 1993. As an example, this is some of the linguistic markup which represents the title of the sample document shown earlier.

macromolecular deformation (NG); macromolecular deformation model (NG); deformation (NG); deformation model (NG); model (NG); viscoelastic flow (NG); viscoelastic flow effects (NGS); flow (NG); flow effects (NGS); effects (NGS); polymer (NG); polymer melts (NGS); melts (NGS)

It is clear that the markup is far from sophisticated, and is very much a small variation on a simple sequence-based representation. Nevertheless, it is fairly accurate in so much as well over 90% of what are claimed to be noun groups can be interpreted as such. One very useful by-product of using a linguistically based representation is that IR can help in linguistic tasks such as terminological collection. I shall present some examples of diagnostic units which are highly associated with descriptors later.

5 Predicting from sophisticated representations

In what follows, we shall compare the relative performance of a term weighting model with the single term model as we vary the sophistication of representation.

Proportional assignment (Lewis 1992b) is used to assign the descriptors from statistical measurements of their appropriateness. This method ensures that roughly the same number of assignments of particular descriptors are made as are actually made in the test corpus. The strategy is simply to assign descriptor d to the N documents which score highest for this descriptor, where N is chosen in proportion to the occurrence of d in the training corpus. For term weighting models, the score is simply the combined weight of the document; for the single term model, the score is $\sup_{w \in \mathcal{R}(x)} P(d|w)$. The *Rule Based* assignment strategy applies only to the single term model and the rule $w \longrightarrow d$ is included just in case $P(d|w) > 1 - \varepsilon$.

Figure 1 shows a few of the rules. All of these entries share the property that $P(d|w) > 0.8$. They were selected at random from the 85,500 associations which were found.

5.1 Representations and models

Five paradigms of representation of documents will be compared, and two term appropriateness models will be compared. This gives us ten combinations. The first representation paradigm is a baseline one: represent documents as the set of the words contained in them. The second paradigm is to represent documents according to word sequences, and the third is to apply a noun-group and adjective-group recogniser. The fourth and fifth representation modes consider representing documents by only their most specific diagnostic units. For example, if the sequence "thermoplastic elastomer compounds"

polymer materials Research/NG;	→	DATA
EEC legislation/NGS;	→	LEGISLATION
venture partners/NGS;	→	JOINT VENTURE
Bergen op/NP	→	PLASTIC
sheet lines/NGS	→	COMPANY
railroad/NG	→	COMPANY
injection moulding facility/NG	→	PLASTIC
PHENOLPHTHALEIN/NP	→	DATA
unsaturated polyester composites/NGS	→	THERMOSET
thermoplastic elastomer compounds/NGS	→	RUBBER
properties features/NGS	→	PLASTIC
fiber Glass/NG	→	GLASS FIBRE REINFORCED PLASTIC
comparative performance/NG	→	DATA
automotive hose/NGS	→	RUBBER
Bitruder/NP	→	EXTRUDER
worldwide tyre/NG	→	COMPANIES
Victrex polyethersulphone/NP	→	COMPANIES
PS melts/NGS	→	PLASTIC
viscoelastic characteristics/NGS	→	VISCOELASTIC PROPERTIES
plastics waste/NG	→	RECYCLING
lattice relaxation/NG	→	NUCLEAR MAGNETIC RESONANCE
fatigue crack propagation/NG	→	MECHANICAL PROPERTIES
unidirectional composites/NGS	→	REINFORCED PLASTIC
Flory Huggins interaction/NG	→	TECHNICAL

Figure 1: This figure shows some probably correct rules for the RAPRA corpus. In all, there are over 85,000 such rules.

appeared in the abstract, then ordinarily this would include the sequence "elastomer compounds", which would be included in the representation. The results of section 3.2 might encourage us to believe that representing a document by only its most specific diagnostic units will improve performance (or, at least, precision). Consequently, a sequence of words is defined to be *most specific* if (a) it is a diagnostic unit and (b) it is not properly contained in a token of any other diagnostic unit present in the document.[2]

The noun-groups are found by performing a simple parse of the documents as described above, and identifying likely noun groups of length 3 or less. The contingency table of diagnostic units verses manually assigned descriptors on a training corpus of 200,000 documents was collected, and this was used as the basis for two term appropriateness models. Probabilities were estimated by adding a constant (usually 0.02 was found fairly optimal) to each cell, and directly estimating from these slightly adjusted counts.

The 50,000 most frequent diagnostic unit types were chosen, and terms which appeared in more than 10% of documents were discarded.

6 Results

The results of the experiments on the RAPRA corpus are presented below.[3]

Despite the peculiarities of the corpus, the message is clear. The result that the standard model fares no better on word sequence sets than on word sets is repeated, and it is clear that the Single Term model fares much better than the Logit model on this data set. However, what is most interesting is that the Single Term models fares significantly better on the more sophisticated sequence based representations of the document than on the simpler word based representation. There is, however, no significant advantage identified by parsing the corpus into noun-groups over simply considering all word sequences. The recall scores for the rule-based tagging strategy show that the improved performance of the sequence based representations can be explained by

[2] If "elastomer compounds" appeared separately in the document from "thermoplastic elastomer compounds", then both of these sequences would be represented in the experiments reported here.

[3] All recall and precision scores are *microaveraged* (Lewis 1992c); they are the expected probability of assigning or recalling correctly per tagging decision. The training set was a set of 200,000 abstracts, and the separate test set had 10,000 abstracts. The experiments looked at only the 520 most common descriptors. In the table, TW means that a term-weighting model was used, while ST means that the single term model was used. 'Word' means the representation was a wordset, 'Seq', the set of all sequences, and 'NG' the set of groups derived from the grammar. For the sequence representations, either all the possible sequences or groups were used (denoted by 'all'), or just the most specific ones were used (denoted by 'spec').

the presence of many more "good" descriptor indicators.

Assignment	Model	Repn	Prec	Rec
Prop.	TW	Word	33%	32%
Prop.	TW	Seq all	32%	34%
Prop.	TW	Seq spec	33%	34%
Prop.	TW	NG all	31%	36%
Prop.	TW	NG spec	32%	32%
Prop.	ST	Word	54%	48%
Prop.	ST	Seq all	57%	55%
Prop.	ST	Seq spec	55%	55%
Prop.	ST	NG	56%	60%
Rule $\varepsilon = .2$	ST	Word	83%	7%
Rule $\varepsilon = .2$	ST	Seq all	77%	42%
Rule $\varepsilon = .2$	ST	Seq spec	80%	40%
Rule $\varepsilon = .2$	ST	NG all	82%	42%
Rule $\varepsilon = .2$	ST	NG spec	84%	37%

7 Conclusion

The significant theoretical result is that as the sophistication of the representation of abstracts is increased, the performance of the single term model improves, while the performance of the term weighting models does not improve significantly. This has been a fairly universal experience among researchers working within the term weighting classification paradigm.

Although there is a very marginally significant improvement from using linguistically sophisticated representations over simple sequence representations if all of the sequences are represented, this largely (though not entirely) disappears when only most specific sequences are considered, so it might be a result of the effects discussed in section 3.2.

The rule based assignment strategy exploits the Single Term model's estimates, and also performs much better on word sequence representations than on word set representations. This assignment strategy is promising because it can exploit more sophisticated representations well, has a sound theory behind it, and will assign descriptors only where it has enough information to do so. Some of the descriptors in the RAPRA corpus, for example, are only ever assigned from the entire article from which the abstract is taken, so no assignment strategy will ever do well on these. On the other hand this model also shows promise that IR techniques might be applied to help infer linguistic resources such as term banks from large classified corpora.

The next stage is to add more sophisticated linguistic annotation to corpora, and to trawl for rules in boolean combinations of descriptors, thus addressing the results of section 3.2. In this way this work can be considered similar in spirit to that undertaken by Apte et al (1994), but differs in the forms of representation which are being considered for documents.

References

Adams, E. (1975) *The Logic of Conditionals: an application of probability to deductive logic* Reidel.

Apte, C, F. Demerau & S. Weiss (1994) Towards Language Independent Automated Learning of Text Categorization Methods. *the proceeding of the Seventeenth ACM-SIGIR Conference on Information Retrieval.* 23–30, DCU, Dublin.

Buckley, C. (1993) The Importance of Proper Weighting Methods. *ARPA Workshop on Human Language Technology.*

Church, K. (1988) A stochastic parts program and noun phrase parser for unrestricted text. In *Second conference on applied NLP,* pp 136–43.

Church, K., W. Gale, P. Hanks & D. Hindle (1989) Parsing, Word Associations and Typical Predicate-Argument Relations. In *International Parsing Technologies Workshop.* CMU, Pittsburgh.

Fagan, J. (1987) *Experiments in Automatic Phrase Indexing for Document Retrieval: Comparison of Syntactic and Non-Syntactic Methods.* PhD Thesis. Cornell University, Dept. of Computer Science.

Finch, S. P. & N. Chater (1991) A Hybrid Approach to the Automatic Learning of Linguistic Categories. *Artificial Intelligence and Simulated Behaviour Quarterly.* **78** 16–24.

Finch, S. (1993) *Finding Structure in Language.* Ph.D. thesis, Centre for Cognitive Science, University of Edinburgh, Edinburgh.

Fuhr, N. (1989) Models for retrieval with probabilistic indexing. *Information processing and management.* **25**(1): 55–72.

Fuhr, N. & Buckley, C (1993) Optimizing Document Indexing and Search Term Weighting Based on Probabilistic Models First TREC Conference.

Hindle, D. (1990) Noun Classification from Predicate-Argument Structures. In *Proceedings of the 22nd meeting of the Association of Computational Linguistics.* 268–75.

Jacobs, P. & Rau, L. (1990) SCISOR: Extracting Information from On-line News *Correspondence of the ACM* **33** 11 88–97

Kupiec, J. (1992) Robust part-of-speech tagging using a hidden Markov model. *Computer Speech and Language,* **6**:3 225–42.

Lewis, D. (1991) Evaluating text categorisation. In *Speech and natural language workshop.* pp 136–143.

Lewis, D. (1992a) *Representation and learning in information retrieval.* Ph.D. thesis, Computer Science Dept., Univ. Mass., Amherst, Ma.

Lewis, D. (1992b) An Evaluation of Phrasal and Clustered Representations on a Text categorization problem. *Proceedings of SIGIR 92.*

Lewis, D. (1992c) Feature selection and feature extraction for text categorization. In *Speech and Natural Language: Proceedings of a Workshop held at Harrimn, NY.* pp 212–217.

Lewis, D. & K. Sparck-Jones (1993) Natural language processing for information retrieval *University of Cambridge Technical report 307,* Cambridge.

Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* Morgan Kaufmann, San Mateo, Ca.

van Rijsbergen, C. J. (1979) *Information retrieval.* Butterworths, London.

Sacks-Davis, R. (1990) Using Syntactic Analysis in a Document Retrieval System that Uses Signature Files. *ACM SIGIR-90.*

Salton, G. & McGill, M. J. (1983) *Introduction to modern information retrieval.* McGraw-Hill, NY.

Salton, G. & C. Buckley (1988) Term Weighting Approaches in Automatic Text Retrieval *Information Processing and Management* **24** 5 513–23

Zadeh, L. (1965) Fuzzy Sets *Information and control, bf 8* 338–53.