Journal of Documentation

A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL

KAREN SPARCK JONES

A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL

KAREN SPARCK JONES
University of Cambridge Computer Laboratory

The exhaustivity of document descriptions and the specificity of index terms are usually regarded as independent. It is suggested that specificity should be interpreted statistically, as a function of term use rather than of term meaning. The effects on retrieval of variations in term specificity are examined, experiments with three test collections showing in particular that frequently-occurring terms are required for good overall performance. It is argued that terms should be weighted according to collection frequency, so that matches on less frequent, more specific, terms are of greater value than matches on frequent terms. Results for the test collections show that considerable improvements in performance are obtained with this very simple procedure.

EXHAUSTIVITY AND SPECIFICITY

WE ARE FAMILIAR with the notions of exhaustivity and specificity: exhaustivity is a property of index descriptions, and specificity one of index terms. They are most clearly illustrated by a simple keyword or descriptor system. In this case the exhaustivity of a document description is the coverage of its various topics given by the terms assigned to it; and the specificity of an individual term is the level of detail at which a given concept is represented.

These features of a document retrieval system have been discussed by Cleverdon *et al.*[4] and Lancaster,[6] for example, and the effects of variation in either have been noted. For instance, if the exhaustivity of a document description is increased by the assignment of more terms, when the number of terms in the indexing vocabulary is constant, the chance of the document matching a request is increased. The idea of an optimum level of indexing exhaustivity for a given document collection then follows: the average number of descriptors per document should be adjusted so that, hopefully, the chances of requests matching relevant documents are maximized, while too many false drops are avoided. Exhaustivity obviously applies to requests too, and one function of a search strategy is to vary request exhaustivity. I shall be mainly concerned here, however, with document descriptions.

Specificity as characterized above is a semantic property of index terms: a term is more or less specific as its meaning is more or less detailed and precise. This is a natural view for anyone concerned with the construction of

an entire indexing vocabulary. Some decision has to be made about the discriminating power of individual terms in addition to their descriptive propriety. For example, the index term 'beverage' may be as properly used for documents about tea, coffee, and cocoa as the terms 'tea', 'coffee', and 'cocoa'. Whether the more general term 'beverage' only is incorporated in the vocabulary, or whether 'tea', 'coffee', and 'cocoa' are adopted, depends on judgements about the retrieval utility of distinctions between documents made by the latter but not the former. It is also predicted that the more general term would be applied to more documents than the separate terms 'tea', 'coffee', and 'cocoa', so the less specific term would have a larger collection distribution than the more specific ones.

It is of course assumed here that such choices when a vocabulary is constructed are exclusive: we may either have 'beverage' or 'tea', 'coffee', and 'cocoa'. What happens if we have all four terms is a different matter. We may then either interpret 'beverage' to mean 'other beverages' or explicitly treat it as a related broader term. I shall, however, disregard these alternatives here.

In setting up an index vocabulary the specificity of index terms is looked at from one point of view: we are concerned with the probable effects on document description, and hence retrieval, of choosing particular terms, or rather of adopting a certain set of terms. For our decisions will in part be influenced by relations between terms, and how the set of chosen terms will collectively characterize the set of documents. But throughout we assume some level of indexing exhaustivity. We are concerned with obtaining an effective vocabulary for a collection of documents of some broadly known subject matter and size, where a given level of indexing exhaustivity is believed to be sufficient to represent the content of individual documents adequately, and distinguish one document from another.

Index term specificity must, however, be looked at from another point of view. What happens when a given index vocabulary is actually used? We predict when we opt for 'beverage', for example, that it will be used more than 'cocoa'. But we do not have much idea of how many documents there will be to which 'beverage' may appropriately be assigned. This is not simply determined even when some level of exhaustivity is assumed. There will be some documents which cry out for 'beverage', so to speak, and we may have some idea of what proportion of the collection this is likely to be. There will also be documents to which 'beverage' cannot justifiably be assigned, and this proportion may also be estimated. But there is unfortunately liable to be some number of documents to which 'beverage' may or may not be assigned, in either case quite plausibly. In general, therefore, the actual use of a descriptor may diverge considerably from the predicted use. The proportions of a collection to which a term does and does not belong can only be estimated very roughly; and there may be enough intermediate documents for the way the term is assigned to these to affect its overall

distribution considerably. Over a long period the character of the collection as a whole may also change, with further effects on term distribution.

This is where the level of exhaustivity of description matters. As a collection grows maintaining a certain level of exhaustivity may mean that the descriptions of different documents are not sufficiently distinguished, while some terms are very heavily used. More generally, great variation in term distribution is likely to appear. It may thus be the case that a particular term becomes less effective as a means of retrieval, whatever its actual meaning. This is because it is not discriminating. It may be properly assigned to documents, in the sense that their content justifies the assignment; but it may no longer be sufficiently useful in itself as a device for distinguishing the typically small class of documents relevant to a request from the remainder of the collection. A frequently used term thus functions in retrieval as a non-specific term, even though its meaning may be quite specific in the ordinary sense.

STATISTICAL SPECIFICITY

It is not enough, in other words, to think of index term specificity solely in setting up an index vocabulary, as having to do with accuracy of concept representation. We should think of specificity as a function of term use. It should be interpreted as a statistical rather than semantic property of index terms. In general we may expect vaguer terms to be used more often, but the behaviour of individual terms will be unpredictable. We can thus redefine exhaustivity and specificity for simple term systems: the exhaustivity of a document description is the number of terms it contains, and the specificity of a term is the number of documents to which it pertains. The relation between the two is then clear, and we can see, for instance, that a change in the exhaustivity of descriptions will affect term specificity: if descriptions are longer, terms will be used more often. This is inevitable for a controlled vocabulary, but also applies if extracted keywords are used, particularly in stem form. The incidence of words new to the keyword vocabulary does not simply parallel the number of documents indexed, and the extraction of more keywords per document is more likely to increase the frequency of current keywords than to generate new ones.

Once this statistical interpretation of specificity, and the relation between it and exhaustivity, are recognized, it is natural to attempt a more formal approach to seeking an optimum level of specificity in a vocabulary and an optimum level of exhaustivity in indexing, for a given collection. Within the broad limits imposed by having sensible terms, i.e. ones which can be reached from requests and applied to documents, we may try to set up a vocabulary with the statistical properties which are hopefully optimal for retrieval. Purely formal calculations may suggest the correct number of terms, and of terms per document, for a certain degree of document discrimination. Work on these lines has been done by Zunde and Slamecka,[10]

for instance. More informally, the suggestion that descriptors should be designed to have approximately the same distribution, made by Salton for example, is motivated by respect for the retrieval effects of purely statistical features of term use.

Unfortunately abstract calculations do not select actual terms. Nor are document collections static. More importantly, it is difficult to control requests. One may characterize documents with a view to distinguishing them nicely and then find that users do not provide requests utilizing these distinctions. We may therefore be forced to accept a de facto non-optimal situation with terms of varying specificity and at least some disagreeably non-specific terms. There will be some terms which, whatever the original intention, retrieve a large number of documents, of which only a small proportion can be expected to be relevant to a request. Such terms are on the whole more of a nuisance than rare, over-specific terms which fail to retrieve documents.

These features of term behaviour can be illustrated by examples from three well-known test collections, obtained from the Aslib Cranfield, INSPEC, and College of Librarianship Wales projects. In fact in these the vocabulary consists of extracted keyword stems, which may be expected to show more variation than controlled terms. But there is no reason to suppose that the situation is essentially different. Full descriptions of the collections are given in Cleverdon *et al*,[4] Aitchison *et al*,[1] and Keen (forthcoming). Relevant characteristics of the collections are given in Section A of Table 1. The INSPEC Collection, for instance, has 541 documents in-

TABLE 1

		Cranfield	*INSPEC*	*Keen*
A.	Number of documents	200	541	797
	Number of terms	712	1,341	939
	Number of terms per document	32	12·2	7·9
	Number of documents per term	9	4·9	6·1
B.	Number of requests	42	97	63
	Number of terms represented	166	248	183
	Number of terms per request	6·9	5·6	5·3
	Number of documents per request term	31·6	11·5	44·8
C.	Number of retrieving terms per request	5	3·2	3·3
	Number of retrieving terms per document	1·8	1·2	1·2
	Number of retrieving terms per relevant document	3·6	2	1·8
D.	Number of frequent terms	96	73	50
	Number of frequent terms per request	4	2·5	2·3

dexed by 1,341 terms. In all the collections, there are some very frequently occurring terms: for example in the Cranfield collection, one term occurs in 144 out of 200 documents; in the INSPEC one term occurs in 112 out of 541, and in the Keen collection one term occurs in 199 out of 797 documents. The terms concerned do not necessarily represent concepts central to the subject areas of the collections, and they are not always general terms. In the Keen collection, which is about information science, the most frequent term is 'index–', and other frequent ones include 'librar–', 'inform–', and 'comput–'. In the INSPEC collection the most frequent is 'theor–', followed by 'measur–' and 'method–'. And in the Cranfield collection the most frequent is 'flow–', followed by 'pressur–', 'distribut–' and 'bound–' (boundary). The rarer terms are a fine mixed bag including 'purchas–', and 'xerograph–' for Keen, 'parallel–' and 'silver–' for INSPEC, and 'logarithm–' and 'seri–' (series) for Cranfield.

SPECIFICITY AND MATCHING

How should one cope with variable term specificity, and especially with insufficiently specific terms, when these occur in requests? The untoward effects of frequent term use can in principle be dealt with very naturally, through term combinations. For instance, though the three terms 'bound–', 'layer–', and 'flow–' occur in 73, 62, and 144 documents each in the Cranfield collection, there are only fifty documents indexed by all three terms together. Relying on term conjunction is quite straightforward. It is in particular a way of overcoming the untoward consequences of the fact that requests tend to be formulated in better known, and hence generally more frequent, terms. It is unfortunate, but not surprising, that requests tend to be presented in terms with an average frequency much above that for the indexing vocabulary as a whole. This holds for all three test collections, as appears in Section B of Table 1. For the Cranfield collection, for example, the average number of postings for the terms in the vocabulary is nine, while the average for the terms used in the requests is 31·6; for Keen the figures are 6·1 and 44·8.

But relying on term combination to reduce false drops is well-known to be risky. It is true that the more terms in common between a document and a request, the more likely it is that the document is relevant to the request. Unfortunately, it just happens to be difficult to match term conjunctions. This is well exhibited by the term matching behaviour of the three collections, as shown in Section C of Table 1. The average number of starting terms per request ranges from 5·3 for Keen to 6·9 for Cranfield. But the average number of retrieving terms per request, i.e. the average of the highest matching scores, ranges from 3·2 to 5·0. More importantly, the average number of matching terms for the relevant documents retrieved ranges from only 1·8 for Keen to 3·6 for Cranfield, though fortunately

the average for all documents retrieved, which are predominantly non-relevant, ranges from a mere 1·2 to 1·8.

Clearly, one solution to this problem is to provide for more matching terms in some way. This may be achieved either by providing alternative substitutes for given terms, through a classification; or by increasing the exhaustivity of document or request specifications, say by adding statistically associated terms. But either approach involves effort, perhaps considerable effort, since the sets of terms related to individual terms must be identified. The question naturally arises as to whether better use of existing term descriptions can be made which does not involve such effort.

As very frequently occurring terms are responsible for noise in retrieval, one possible course is simply to remove them from requests. The fact that this will reduce the number of terms available for conjoined matching may be offset by the fact that fewer non-relevant documents will be retrieved. Unfortunately, while frequent terms cause noise, they are also required for reasonably high recall. For all three test collections, the deletion of very frequent terms by the application of a suitable threshold leads to a decline in overall performance. For the INSPEC collection, for example, the threshold was set to delete terms occurring in twenty or more documents, so that seventy-three terms out of the total vocabulary of 1,341 were removed. The effect in retrieval performance is illustrated by the recall/precision graph of Figure 1 for the Cranfield collection. Matching is by simple term

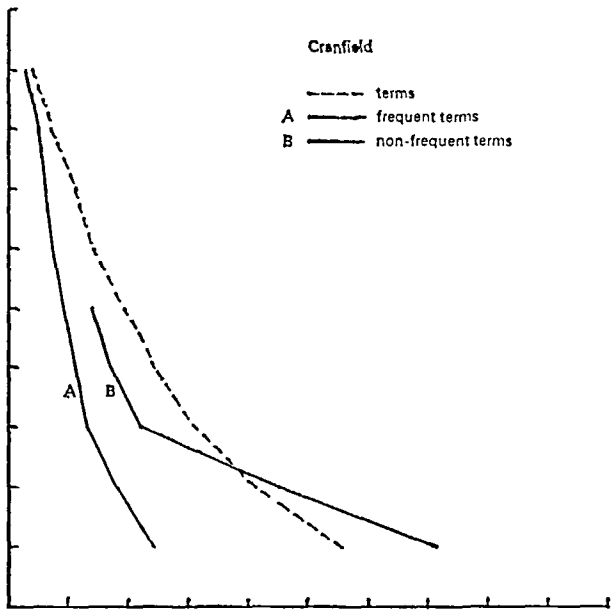

FIG. 1

co-ordination levels, and averaging over the set of requests is by straightforward average of numbers. Precision at ten standard recall values is then interpolated. The same relationship between full term matching and this restricted matching with non-frequent terms only is exhibited by the other collections: the recall ceiling is lowered by at least 30%, and indeed for the Keen collection is reduced from 75% to 25%, though precision is maintained.

Inspection of the requests shows why this result is obtained. Not merely is request term frequency much above average collection frequency; the comparatively small number of very frequent terms plays a large part in request formulation. 'Flow–' for example, appears in twelve Cranfield requests out of forty-two, and in general for all three collections about half the terms in a request are very frequent ones, as shown in Section D of Table 1. Throwing very frequent terms away is throwing the baby out with the bath water, since they are required for the retrieval of many relevant documents. The combination of non-frequent terms is discriminating, but no more than that of frequent and non-frequent terms. The value of the non-frequent terms is clearly seen, on the other hand, when matching using frequent terms only is compared with full matching, also shown in Figure 1. Matching levels for total and relevant documents are nearly as high as for all terms, but the non-frequent terms in the latter raise the relevant matching level about 1.

These features of term retrieval suggest that to improve on the initial full term performance we need to exploit the good features of very frequent and non-frequent terms, while minimizing their bad ones. We should allow some merit in frequent term matches, while allowing rather more in non-frequent ones. In any case we wish to maximize the number of matching terms.

WEIGHTING BY SPECIFICITY

This clearly suggests a weighting scheme. In normal term co-ordination matches, if a request and document have a frequent term in common, this counts for as much as a non-frequent one; so if a request and document share three common terms, the document is retrieved at the same level as another one sharing three rare terms with the request. But it seems we should treat matches on non-frequent terms as more valuable than ones on frequent terms, without disregarding the latter altogether. The natural solution is to correlate a term's matching value with its collection frequency. At this stage the division of terms into frequent and non-frequent is arbitrary and probably not optimal: the elegant and almost certainly better approach is to relate matching value more closely to relative frequency. The appropriate way of doing this is suggested by the term distribution curve for the vocabulary, which has the familiar Zipf shape. Let $f(n) = m$ such that $2^{m-1} < n \leqslant 2^m$. Then where there are N documents in the collection, the weight of a term which

occurs n times is $f(N) - f(n) + 1$. For the Cranfield collection with 200 documents, for example, this means that a term occurring ninety times has weight 2, while one occurring three times has weight 7.

The matching value of a term is thus correlated with its specificity and the retrieval level of a document is determined by the sum of the values of its matching terms. Simple co-ordination levels are replaced by a more sophisticated quasi-ranking. The effect can be illustrated by the different retrieval levels at which two documents matching a request on the same number of relatively frequent and relatively non-frequent terms respectively. With the Cranfield range of values, a document matching on two terms with frequencies 15 and 43 will be retrieved at level $5 + 3 = 8$, while one matching on terms with frequencies 3 and 7 will be retrieved at level $7 + 6 = 13$. Clearly, as the range of levels is 'stretched', more discrimination is possible.

The idea of term weighting is not new. But it is typically related to the presumed importance of a term with respect to a document in itself. For instance, if a document is mainly about paint and only mentions varnish in passing, we may utilize some simple weighting scale to assign a weight of 2 to the term 'paint' and 1 to 'varnish'. More informally, in putting a request, we may state that during searching term x must be retained, but term y may be dropped. More systematic weighting on a statistical base may be adopted if the necessary information is available. If the actual frequency of occurrence of terms in a document (or abstract) is known, this may be used to generate weights. Artandi and Wolfe[2] report the use of frequency to select a weight from a three-point scale, while Salton[7] more wholeheartedly uses the frequency of occurrence as a weight. In a range of experiments Salton has demonstrated that weighting terms in this way leads to a noticeable improvement in performance over that obtained for unweighted terms.

Weighting by collection frequency as opposed to document frequency is quite different. It places greater emphasis on the value of a term as a means of distinguishing one document from another than on its value as an indication of the content of the document itself. The relation between the two forms of weighting is not obvious. In some cases a term may be common in a document and rare in the collection, so that it would be heavily weighted in both schemes. But the reverse may also apply. It is really that the emphasis is on different properties of terms.

The treatment of term collection frequency in connection with term matching does not seem to have been systematically investigated. The effect of term frequency on statistical associations has been studied, for example by Lesk, but this is a different matter. The fact that a given term is likely to retrieve a large number of documents may be informally exploited in setting up searches, in particular in the context of on-line retrieval as described by Borko[5] for example. More whole-hearted approaches are prob-

ably hampered by the lack of the necessary information. Such a procedure as the one described is also much more suited to automatic than manual searching. It is of interest, therefore, that term frequencies have been exploited in the general manner indicated within an operational interactive retrieval system for internal reports implemented at A. D. Little (Curtice and Jones[5]). In this system indexing keywords are extracted automatically from text, and the weighting is therefore associated with a changing vocabulary and collection. However, no systematic experiments are reported.

EXPERIMENTAL RESULTS

The term weighting system described was tried on the three collections. As noted, these are very different in character, with different sizes of vocabulary, document description, and request specification, as indicated in Table 1. In all cases, however, matching with term weighting led to a substantial improvement in performance over simple term matching. The results presented in the form mentioned earlier, are given in Figure 2. A simple significance test based on the difference in area enclosed by the curves shows that the improvement given by weighted terms is fully significant, the difference being well above the required minimum.

These results are of interest for two reasons. All three collections have been used for a whole range of experiments with different index languages,

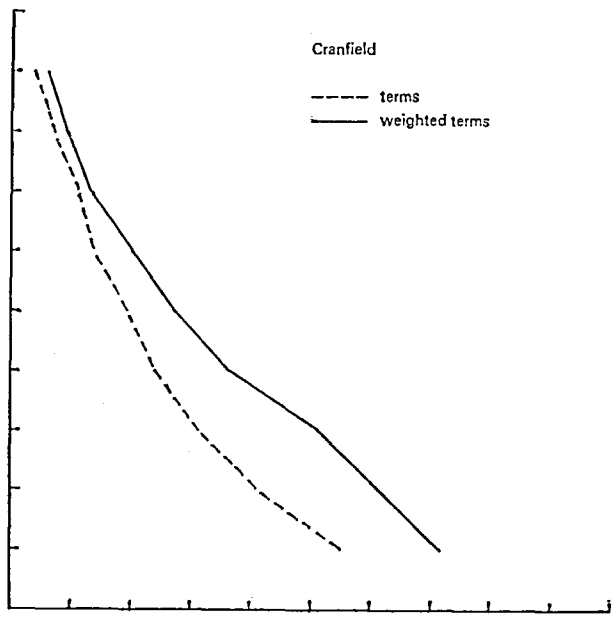

FIG. 2*a*

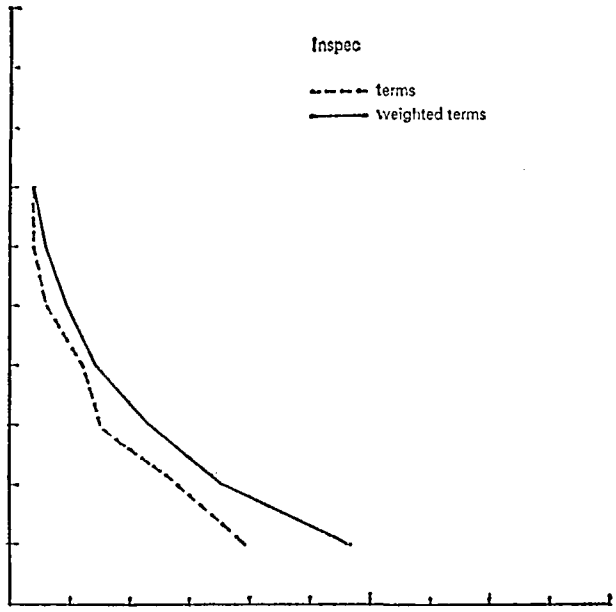

FIG. 2*b*

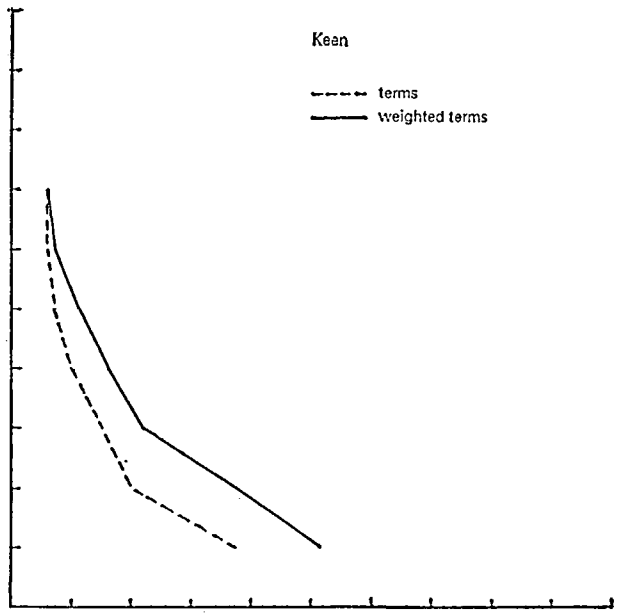

FIG. 2*c*

search techniques, and so on: see Cleverdon *et al*,[4] Salton,[7] Salton and Lesk,[8] and Sparck Jones;[9] Aitchison *et al*[1] and Keen (forthcoming). The performance improvement obtained here nevertheless represents as good an improvement over simple unweighted keyword stem matching as has been obtained by any other means, including carefully constructed thesauri: Salton's iterative search methods are not comparable. The details of the way these experimental results are presented varies, so rigorous comparisons are impossible: but the general picture is clear. Indeed, in so far as anything can be called a solid result in information retrieval research, this is one. The second point about the present results is that the improvement in performance is obtained by extremely simple means. It is compatible with an initially plain method of indexing, namely the use of extracted keywords, which may be reduced to stems automatically; it is readily implemented given an automatic term-matching procedure, since all that is required is a term frequency list and this is easily obtained; and it has the merit that the weight assigned to terms is naturally adjusted to follow the growth of and changes in a collection. Experiments with very much larger collections than those used here are clearly desirable; they will hopefully not be long delayed.

REFERENCES

1. AITCHISON, T. M., HALL, A. M., LAVELLE, K. H., *and* TRACY, J. M. *Comparative evaluation of index languages, Part II; Results*, Project INSPEC, Institute of Electrical Engineers, London, 1970.
2. ARTANDI, S. *and* WOLF, E. H. The effectiveness of automatically-generated weights and links, *American Documentation*, vol. 20, 1969, p. 198–202.
3. BORKO, H. Interactive document storage and retrieval systems—design concepts, *Mechanised information storage, retrieval and dissemination* (ed. Samuleson), Amsterdam, North-Holland, 1968, p. 591–99.
4. CLEVERDON, C. W., MILLS, J., *and* KEEN, E. M. *Factors determining the performance of indexing systems*, 2 vols. Cranfield, 1966.
5. CURTICE, R. M. *and* JONES, P. E. An operational interactive retrieval system. Cambridge, Mass., Arthur D. Little Inc., 1969.
6. LANCASTER, F. W., *Information retrieval systems: characteristics, testing and evaluation*, New York, Wiley, 1968.
7. SALTON, G. *Automatic information organization and retrieval*, New York, McGraw-Hill, 1968.
8. SALTON, G. *and* LESK, M. E. Computer evaluation of indexing and text processing, *Journal of the ACM*, vol. 15, 1968, p. 8–36.
9. SPARCK JONES, K. *Automatic keyword classification for information retrieval*, London, Butterworths, 1971.
10. ZUNDE, P. *and* SLAMECKA, V. Distribution of indexing terms for maximum efficiency of information transmission, *American Documentation*, vol. 18, 1967, p. 104–8.

This article has been cited by:

1. Yogesh Gupta, Ashish Saini, A.K. Saxena. 2015. A new fuzzy logic based ranking function for efficient Information Retrieval system. *Expert Systems with Applications* **42**, 1223-1234. [CrossRef]
2. J.D. Echeverry-Correa, J. Ferreiros-López, A. Coucheiro-Limeres, R. Córdoba, J.M. Montero. 2015. Topic identification techniques applied to dynamic language model adaptation for automatic speech recognition. *Expert Systems with Applications* **42**, 101-112. [CrossRef]
3. J. Bernabé-Moreno, A. Tejeda-Lorente, C. Porcel, E. Herrera-Viedma. 2014. A new model to quantify the impact of a topic in a location over time with Social Media. *Expert Systems with Applications* . [CrossRef]
4. Howard D. White. 2014. Co-cited author retrieval and relevance theory: examples from the humanities. *Scientometrics* . [CrossRef]
5. D. Kiela, Y. Guo, U. Stenius, A. Korhonen. 2014. Unsupervised discovery of information structure in biomedical documents. *Bioinformatics* . [CrossRef]
6. W. Yan, H. Liu, C. Zanni-Merk, D. Cavallucci. 2014. IngeniousTRIZ: An automatic ontology-based system for solving inventive problems. *Knowledge-Based Systems* . [CrossRef]
7. Yuanben Zhang, Yu Huang, Kun Fu, Jun Song, Xiang Qi. 2014. TextInsight: A new text visualization system based on entropy and GMap. *Journal of Electronics (China)* **31**, 453-464. [CrossRef]
8. Valentin Tablan, Kalina Bontcheva, Ian Roberts, Hamish Cunningham. 2014. Mímir: An open-source semantic search framework for interactive information seeking and discovery. *Web Semantics: Science, Services and Agents on the World Wide Web* . [CrossRef]
9. Frederik C. Schadd, Nico Roos. 2014. Word-Sense Disambiguation for Ontology Mapping: Concept Disambiguation using Virtual Documents and Information Retrieval Techniques. *Journal on Data Semantics* . [CrossRef]
10. Bibliography 309-342. [CrossRef]
11. Makarand Tapaswi, Martin Bäuml, Rainer Stiefelhagen. 2014. Aligning plot synopses to videos for story-based retrieval. *International Journal of Multimedia Information Retrieval* . [CrossRef]
12. Shahid Alam, Ibrahim Sogukpinar, Issa Traore, R. Nigel Horspool. 2014. Sliding window and control flow weight for metamorphic malware detection. *Journal of Computer Virology and Hacking Techniques* . [CrossRef]
13. Y. Ni, S. Kennebeck, J. W. Dexheimer, C. M. McAneney, H. Tang, T. Lingren, Q. Li, H. Zhai, I. Solti. 2014. Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department. *Journal of the American Medical Informatics Association* . [CrossRef]
14. Hankz Hankui Zhuo, Qiang Yang. 2014. Action-model acquisition for planning via transfer learning. *Artificial Intelligence* **212**, 80-103. [CrossRef]
15. Yanhui Gu, Zhenglu Yang, Guandong Xu, Miyuki Nakano, Masashi Toyoda, Masaru Kitsuregawa. 2014. Exploration on efficient similar sentences extraction. *World Wide Web* **17**, 595-626. [CrossRef]
16. Emanuela Riviera. 2014. Testing the strength of the normative approach in citation theory through relational bibliometrics: The case of italian sociology. *Journal of the Association for Information Science and Technology* n/a-n/a. [CrossRef]
17. Birger Hjørland. 2014. Classical databases and knowledge organization: A case for boolean retrieval and human decision-making during searches. *Journal of the Association for Information Science and Technology* n/a-n/a. [CrossRef]

18. Yu Liu, Weijia Li, Zhen Huang, Qiang Fang. 2014. A fast method based on multiple clustering for name disambiguation in bibliographic citations. *Journal of the Association for Information Science and Technology* n/a-n/a. [CrossRef]
19. Peng Tang, Mingbo Zhao, Tommy W.S. Chow. 2014. Text style analysis using trace ratio criterion patch alignment embedding. *Neurocomputing* **136**, 201-212. [CrossRef]
20. Matteo Golfarelli, Elisa Turricchia. 2014. A characterization of hierarchical computable distance functions for data warehouse systems. *Decision Support Systems* **62**, 144-157. [CrossRef]
21. Amir H. Razavi, Stan Matwin, Joseph De Koninck, Ray Reza Amini. 2014. Dream sentiment analysis using second order soft co-occurrences (SOSCO) and time course representations. *Journal of Intelligent Information Systems* **42**, 393-413. [CrossRef]
22. Aditya Ramana Rachakonda, Srinath Srinivasa, Sumant Kulkarni, M.S. Srinivasan. 2014. A generic framework and methodology for extracting semantics from co-occurrences. *Data & Knowledge Engineering* . [CrossRef]
23. Zhi-Hong Deng, Kun-Hu Luo, Hong-Liang Yu. 2014. A study of supervised term weighting scheme for sentiment analysis. *Expert Systems with Applications* **41**, 3506-3513. [CrossRef]
24. Jiachen Yang, Keisuke Hotta, Yoshiki Higo, Hiroshi Igaki, Shinji Kusumoto. 2014. Classification model for code clones based on machine learning. *Empirical Software Engineering* . [CrossRef]
25. Chinmayee Annachhatre, Thomas H. Austin, Mark Stamp. 2014. Hidden Markov models for malware classification. *Journal of Computer Virology and Hacking Techniques* . [CrossRef]
26. María José Marín. 2014. Evaluation of five single-word term recognition methods on a legal English corpus. *Corpora* **9**, 83-107. [CrossRef]
27. Abdelbadie Belmouhcine, Mohammed Benkhalifa. 2014. Hardware Implementation of Web Based Arabic Optical Character Recognition Units. *Journal of Emerging Technologies in Web Intelligence* **6**. . [CrossRef]
28. M. Schulz, F. Krause, N. Le Novere, E. Klipp, W. Liebermeister. 2014. Retrieval, alignment, and clustering of computational models based on semantic annotations. *Molecular Systems Biology* **7**, 512-512. [CrossRef]
29. R. S. Forsyth, S. Sharoff. 2014. Document dissimilarity within and across languages: A benchmarking study. *Literary and Linguistic Computing* **29**, 6-22. [CrossRef]
30. İlker Kocabaş, Bekir Taner Dinçer, Bahar Karaoğlan. 2014. A nonparametric term weighting method for information retrieval based on measuring the divergence from independence. *Information Retrieval* **17**, 153-176. [CrossRef]
31. BO JI, YANGDONG YE, YU XIAO. 2014. A COMBINATION WEIGHTING ALGORITHM USING RELATIVE ENTROPY FOR DOCUMENT CLUSTERING. *International Journal of Pattern Recognition and Artificial Intelligence* 1453002. [CrossRef]
32. Christopher Boston, Hui Fang, Sandra Carberry, Hao Wu, Xitong Liu. 2014. Wikimantic: Toward effective disambiguation and expansion of queries. *Data & Knowledge Engineering* **90**, 22-37. [CrossRef]
33. Yen-Liang Chen, Ching-Hao Chuang, Yu-Ting Chiu. 2014. Community detection based on social interactions in a social network. *Journal of the Association for Information Science and Technology* **65**:10.1002/asi.2014.65.issue-3, 539-550. [CrossRef]
34. Stefano Baccianella, Andrea Esuli, Fabrizio Sebastiani. 2014. Feature Selection for Ordinal Text Classification. *Neural Computation* **26**, 557-591. [CrossRef]
35. Bibliography 177-197. [CrossRef]

36. Edward K. F. Dang, Robert W. P. Luk, James Allan. 2014. Beyond bag-of-words: Bigram-enhanced context-dependent term weights. *Journal of the Association for Information Science and Technology* n/a-n/a. [CrossRef]
37. Hollie C. White. 2014. Descriptive Metadata for Scientific Data Repositories: A Comparison of Information Scientist and Scientist Organizing Behaviors. *Journal of Library Metadata* **14**, 24-51. [CrossRef]
38. Jinn-Tsong Tsai. 2014. Optimized weights of document keywords for auto-reply accuracy. *Neurocomputing* **124**, 43-56. [CrossRef]
39. Haitao Yu, Fuji Ren. 2014. Automatic role-explicit query extraction: a divide-and-conquer system leveraging on users' reformulating behaviors. *IEEJ Transactions on Electrical and Electronic Engineering* **9**:10.1002/tee.v9.1, 62-70. [CrossRef]
40. Ángel Felices-Lago, Pedro Ureña Gómez-MorenoFunGramKB term extractor: A tool for building terminological ontologies
from specialised corpora 251-270. [CrossRef]
41. Jun-Ming Chen, Meng-Chang Chen, Yeali S. Sun. 2014. A tag based learning approach to knowledge acquisition for constructing prior knowledge and enhancing student reading comprehension. *Computers & Education* **70**, 256-268. [CrossRef]
42. W. Yan, C. Zanni-Merk, D. Cavallucci, P. Collet. 2014. An ontology-based approach for inventive problem solving. *Engineering Applications of Artificial Intelligence* **27**, 175-190. [CrossRef]
43. Kai Sun, Natalie Buchan, Chris Larminie, Nataša Pržulj. 2014. The integrated disease network. *Integr. Biol.* **6**, 1069-1079. [CrossRef]
44. Irena Pletikosa Cvijikj, Erica Dubach Spiegler, Florian Michahelles. 2013. Evaluation framework for social media brand presence. *Social Network Analysis and Mining* **3**, 1325-1349. [CrossRef]
45. Dong Zhou, Mark Truran, Jianxun Liu, Sanrong Zhang. 2013. Collaborative pseudo-relevance feedback. *Expert Systems with Applications* **40**, 6805-6812. [CrossRef]
46. Alejandro Bellogín, Jun Wang, Pablo Castells. 2013. Bridging memory-based collaborative filtering and text retrieval. *Information Retrieval* **16**, 697-724. [CrossRef]
47. Eva Zangerle, Wolfgang Gassler, Günther Specht. 2013. On the impact of text similarity functions on hashtag recommendations in microblogging environments. *Social Network Analysis and Mining* **3**, 889-898. [CrossRef]
48. Robert M. Losee. 2013. The effect of assigning a metadata or indexing term on document ordering. *Journal of the American Society for Information Science and Technology* **64**:10.1002/asi.2013.64.issue-11, 2191-2200. [CrossRef]
49. Sunyoung Cho, Sooyeong Kwak, Hyeran Byun. 2013. Recognizing human–human interaction activities using visual and textual information. *Pattern Recognition Letters* **34**, 1840-1848. [CrossRef]
50. Sarwat Nizamani, Nasrullah Memon. 2013. CEAI: CCM-based email authorship identification model. *Egyptian Informatics Journal* **14**, 239-249. [CrossRef]
51. ROBERT LAYTON, PAUL WATTERS, RICHARD DAZELEY. 2013. Evaluating authorship distance methods using the positive Silhouette coefficient. *Natural Language Engineering* **19**, 517-535. [CrossRef]
52. Peng Tang, Tommy W.S. Chow. 2013. Recognition of word collocation habits using frequency rank ratio and inter-term intimacy. *Expert Systems with Applications* **40**, 4301-4314. [CrossRef]
53. Xiaozhong Liu, Jinsong Zhang, Chun Guo. 2013. Full-text citation analysis: A new method to enhance scholarly networks. *Journal of the American Society for Information Science and Technology* **64**:10.1002/asi.2013.64.issue-9, 1852-1863. [CrossRef]

54. Arun Lakhotia, Andrew Walenstein, Craig Miles, Anshuman Singh. 2013. VILO: a rapid learning nearest-neighbor classifier for malware triage. *Journal of Computer Virology and Hacking Techniques* **9**, 109-123. [CrossRef]
55. Roland Schäfer, Felix Bildhauer. 2013. Web Corpus Construction. *Synthesis Lectures on Human Language Technologies* **6**, 1-145. [CrossRef]
56. Edgar Gonzàlez, Jordi Turmo. 2013. Unsupervised ensemble minority clustering. *Machine Learning* . [CrossRef]
57. Fuji Ren, Mohammad Golam Sohrab. 2013. Class-indexing-based term weighting for automatic text classification. *Information Sciences* **236**, 109-125. [CrossRef]
58. Mahmoud Elmarhoumy, Mohamed Abdel Fattah, Motoyuki Suzuki, Fuji Ren. 2013. A new modified centroid classifier approach for automatic text classification. *IEEJ Transactions on Electrical and Electronic Engineering* **8**:10.1002/tee.v8.4, 364-370. [CrossRef]
59. Dnyanesh G. Rajpathak. 2013. An ontology based text mining system for knowledge discovery from the diagnosis data in the automotive domain. *Computers in Industry* **64**, 565-580. [CrossRef]
60. Yan Zhu, Jian Yu, Liping Jing. 2013. A novel semi-supervised learning framework with simultaneous text representing. *Knowledge and Information Systems* **34**, 547-562. [CrossRef]
61. C. Carretero-Campos, P. Bernaola-Galván, A.V. Coronado, P. Carpena. 2013. Improving statistical keyword detection in short texts: Entropic and clustering approaches. *Physica A: Statistical Mechanics and its Applications* **392**, 1481-1492. [CrossRef]
62. Logistic Regression and Text Classification 59-84. [CrossRef]
63. Deqing Wang, Junjie Wu, Hui Zhang, Ke Xu, Mengxiang Lin. 2013. Towards enhancing centroid classifier for text classification—A border-instance approach. *Neurocomputing* **101**, 299-308. [CrossRef]
64. Masaki Eto. 2013. Evaluations of context-based co-citation searching. *Scientometrics* **94**, 651-673. [CrossRef]
65. Simson L. Garfinkel. 2013. Digital media triage with bulk data analysis and bulk_extractor. *Computers & Security* **32**, 56-72. [CrossRef]
66. Yunhyong Kim, Seamus Ross. 2013. Closing the loop: Assisting archival appraisal and information retrieval in one sweep. *Proceedings of the American Society for Information Science and Technology* **50**:10.1002/meet.145.v50:1, 1-10. [CrossRef]
67. Lars Olsen, Ulrich Johan Kudahl, Ole Winther, Vladimir Brusic. 2013. Literature classification for semi-automated updating of biological knowledgebases. *BMC Genomics* **14**, S14. [CrossRef]
68. Carlos Arturo Hernández-Gracidas, Luis Enrique Sucar, Manuel Montes-y-Gómez. 2013. Improving image retrieval by using spatial relations. *Multimedia Tools and Applications* **62**, 479-505. [CrossRef]
69. Dirk Thorleuchter, Dirk Van den Poel. 2012. Predicting e-commerce company success by mining the text of its publicly-accessible website. *Expert Systems with Applications* **39**, 13026-13034. [CrossRef]
70. Gerasimos Spanakis, Georgios Siolas, Andreas Stafylopatis. 2012. DoSO: a document self-organizer. *Journal of Intelligent Information Systems* **39**, 577-610. [CrossRef]
71. Dirk Thorleuchter, Dirk Van den Poel. 2012. Improved multilevel security with latent semantic indexing. *Expert Systems with Applications* **39**, 13462-13471. [CrossRef]
72. Chihli Hung, Chih-Fong Tsai, Shin-Yuan Hung, Chang-Jiang Ku. 2012. OGIR: an ontology-based grid information retrieval framework. *Online Information Review* **36**:6, 807-827. [Abstract] [Full Text] [PDF]

73. Antoon Bronselaer, Guy De Tré. 2012. Concept-relational text clustering. *International Journal of Intelligent Systems* **27**:10.1002/int.v27.11, 970-993. [CrossRef]
74. Duen-Ren Liu, Chin-Hui Lai, Ya-Ting Chen. 2012. Document recommendations based on knowledge flows: A hybrid of personalized and group-based approaches. *Journal of the American Society for Information Science and Technology* **63**:10.1002/asi.v63.10, 2100-2117. [CrossRef]
75. J. Driesen, H. Van hamme. 2012. Supervised input space scaling for non-negative matrix factorization. *Signal Processing* **92**, 1864-1874. [CrossRef]
76. SAMI VIRPIOJA, MARI-SANNA PAUKKERI, ABHISHEK TRIPATHI, TIINA LINDH-KNUUTILA, KRISTA LAGUS. 2012. Evaluating vector space models with canonical correlation analysis. *Natural Language Engineering* **18**, 399-436. [CrossRef]
77. Yan Tang, Robert Meersman. 2012. DIY-CDR: an ontology-based, Do-It-Yourself component discoverer and recommender. *Personal and Ubiquitous Computing* **16**, 581-595. [CrossRef]
78. Soner Kara, Özgür Alan, Orkunt Sabuncu, Samet Akpınar, Nihan K. Cicekli, Ferda N. Alpaslan. 2012. An ontology-based retrieval system using semantic indexing. *Information Systems* **37**, 294-305. [CrossRef]
79. Ying Liu, Kun BaiKnowledge-Based Systems Engineering 157-176. [CrossRef]
80. Myungjin Lee, Wooju Kim, Sangun Park. 2012. Searching and ranking method of relevant resources by user intention on the Semantic Web. *Expert Systems with Applications* **39**, 4111-4121. [CrossRef]
81. Roi Blanco, Christina Lioma. 2012. Graph-based term weighting for information retrieval. *Information Retrieval* **15**, 54-92. [CrossRef]
82. Gloria Bordogna, Gabriella Pasi. 2012. A quality driven Hierarchical Data Divisive Soft Clustering for information retrieval. *Knowledge-Based Systems* **26**, 9-19. [CrossRef]
83. Dirk Thorleuchter, Dirk Van den Poel, Anita Prinzie. 2012. Analyzing existing customers' websites to improve the customer acquisition process as well as the profitability prediction in B-to-B marketing. *Expert Systems with Applications* **39**, 2597-2605. [CrossRef]
84. Hiroshi Sugimura, Kazunori Matsumoto. 2012. Development of Knowledge Discovery System from Annotated Time Sereis Data. *IEEJ Transactions on Electronics, Information and Systems* **132**, 592-597. [CrossRef]
85. Byung-Chul So, Jin-Woo Jung. 2011. Implementation of Search Method based on Sequence and Adjacency Relationship of User Query. *Journal of Korean Institute of Intelligent Systems* **21**, 724-729. [CrossRef]
86. J. Zheng, J. Stoyanovich, E. Manduchi, J. Liu, C. J. Stoeckert. 2011. AnnotCompute: annotation-based exploration and meta-analysis of genomics experiments. *Database* **2011**, bar045-bar045. [CrossRef]
87. Sérgio Nunes, Cristina Ribeiro, Gabriel David. 2011. Term weighting based on document revision history. *Journal of the American Society for Information Science and Technology* **62**:10.1002/asi.v62.12, 2471-2478. [CrossRef]
88. Stephen Luther, Donald Berndt, Dezon Finch, Matthew Richardson, Edward Hickling, David Hickam. 2011. Using statistical text mining to supplement the development of an ontology. *Journal of Biomedical Informatics* **44**, S86-S93. [CrossRef]
89. Max L. Wilson. 2011. Search User Interface Design. *Synthesis Lectures on Information Concepts, Retrieval, and Services* **3**, 1-143. [CrossRef]
90. Brian E. Chapman, Sean Lee, Hyunseok Peter Kang, Wendy W. Chapman. 2011. Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm. *Journal of Biomedical Informatics* **44**, 728-737. [CrossRef]

91. Hai Dong, Farookh Khadeer Hussain, Elizabeth Chang. 2011. A framework for discovering and classifying ubiquitous services in digital health ecosystems. *Journal of Computer and System Sciences* **77**, 687-704. [CrossRef]
92. Kazuhiro Seki, Kuniaki Uehara. 2011. Opinionated document retrieval using subjective triggers. *Journal of the American Society for Information Science and Technology* **62**, 861-876. [CrossRef]
93. Yen-Liang Chen, Yu-Ting Chiu. 2011. An IPC-based vector space model for patent retrieval. *Information Processing & Management* **47**, 309-322. [CrossRef]
94. Yan Xu. 2011. A Data-drive Feature Selection Method in Text Categorization. *Journal of Software* **6**. . [CrossRef]
95. Timothy J. HazenTopic Identification 319-356. [CrossRef]
96. Wen Zhang, Taketoshi Yoshida, Xijin Tang. 2011. A comparative study of TF*IDF, LSI and multi-words for text classification. *Expert Systems with Applications* **38**, 2758-2765. [CrossRef]
97. Hai Dong, Farookh Khadeer Hussain, Elizabeth Chang. 2011. A service concept recommendation system for enhancing the dependability of semantic service matchmakers in the service ecosystem environment. *Journal of Network and Computer Applications* **34**, 619-631. [CrossRef]
98. Kook-Yeol Yoon, Sun-Wook Choi, Chong-Ho Lee. 2011. An Approach for Localization Around Indoor Corridors Based on Visual Attention Model. *Journal of Institute of Control, Robotics and Systems* **17**, 93-101. [CrossRef]
99. Paul A. Watters, Stephen McCombie. 2011. A methodology for analyzing the credential marketplace. *Journal of Money Laundering Control* **14**:1, 32-43. [Abstract] [Full Text] [PDF]
100. Taiki Miyanishi, Kazuhiro Seki, Kuniaki Uehara. 2011. Hypothesis Ranking Based on Semantic Event Similarities. *IPSJ Transactions on Bioinformatics* **4**, 9-20. [CrossRef]
101. Hangzai Luo, Jianping Fan, Youjie Zhou. 2011. Multimedia news exploration and retrieval by integrating keywords, relations and visual features. *Multimedia Tools and Applications* **51**, 625-648. [CrossRef]
102. E.K.F. Dang, R.W.P. Luk, J. Allan, K.S. Ho, S.C.F. Chan, K.F.L. Chung, D.L. Lee. 2010. A new context-dependent term weight computed by boost and discount using relevance information. *Journal of the American Society for Information Science and Technology* **61**:10.1002/asi.v61.12, 2514-2530. [CrossRef]
103. Mark Cecchini, Haldun Aytug, Gary J. Koehler, Praveen Pathak. 2010. Making words work: Using financial text as a predictor of financial events. *Decision Support Systems* **50**, 164-175. [CrossRef]
104. Byeong-Chul Kang, Zee-Won Sur, Chulhwan Park, Man-gi Cho. 2010. Document clustering of MEDLINE abstracts based on non-negative matrix factorization using local confidence assessment. *BioChip Journal* **4**, 336-349. [CrossRef]
105. Yong-gang Cao, James J. Cimino, John Ely, Hong Yu. 2010. Automatically extracting information needs from complex clinical questions. *Journal of Biomedical Informatics* **43**, 962-971. [CrossRef]
106. GAËL DIAS, RUMEN MORALIYSKI, JOÃO CORDEIRO, ANTOINE DOUCET, HELENA AHONEN-MYKA. 2010. Automatic discovery of word semantic relations using paraphrase alignment and distributional lexical semantics analysis. *Natural Language Engineering* **16**, 439-467. [CrossRef]
107. Pablo A. D. Castro, Fabrício O. França, Hamilton M. Ferreira, Guilherme Palermo Coelho, Fernando J. Zuben. 2010. Query expansion using an immune-inspired biclustering algorithm. *Natural Computing* **9**, 579-602. [CrossRef]
108. Wen Zhang, Taketoshi Yoshida, Xijin Tang, Qing Wang. 2010. Text clustering using frequent itemsets. *Knowledge-Based Systems* **23**, 379-388. [CrossRef]

109. Howard D. White. 2010. Some new tests of relevance theory in information science. *Scientometrics* **83**, 653-667. [CrossRef]

110. Charles V. Trappey, Amy J.C. Trappey, Chun-Yi Wu. 2010. Clustering patents using non-exhaustive overlaps. *Journal of Systems Science and Systems Engineering* **19**, 162-181. [CrossRef]

111. Stuart Rose, Dave Engel, Nick Cramer, Wendy CowleyAutomatic Keyword Extraction from Individual Documents 1-20. [CrossRef]

112. Valerie Cross, Vishal Bathija. 2010. Automatic ontology creation using adaptation. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* **24**, 127. [CrossRef]

113. Feifan Liu, Yang Liu. 2010. Exploring Correlation Between ROUGE and Human Evaluation on Meeting Summaries. *IEEE Transactions on Audio, Speech, and Language Processing* **18**, 187-196. [CrossRef]

114. Kenichi Mishina, Seiji Tsuchiya, Motoyuki Suzuki, Fuji Ren. 2010. An Improvement of Example-based Emotion Estimation Using Similarity between Sentence and each Corpus. *Journal of Natural Language Processing* **17**, 91-110. [CrossRef]

115. Bernard J. Jansen, Soo Young Rieh. 2010. The seventeen theoretical constructs of information searching and information retrieval. *Journal of the American Society for Information Science and Technology* n/a-n/a. [CrossRef]

116. H. Strobelt, D. Oelke, C. Rohrdantz, A. Stoffel, D.A. Keim, O. Deussen. 2009. Document Cards: A Top Trumps Visualization for Documents. *IEEE Transactions on Visualization and Computer Graphics* **15**, 1145-1152. [CrossRef]

117. Ghassan Kanaan, Riyad Al-Shalabi, Sameh Ghwanmeh, Hamda Al-Ma'adeed. 2009. A comparison of text-classification techniques applied to Arabic text. *Journal of the American Society for Information Science and Technology* **60**:10.1002/asi.v60:9, 1836-1844. [CrossRef]

118. Di Cai. 2009. Determining semantic relatedness through the measurement of discrimination information using Jensen difference. *International Journal of Intelligent Systems* **24**:10.1002/int.v24:5, 477-503. [CrossRef]

119. Man Lan, Chew Lim Tan, Jian Su, Yue Lu. 2009. Supervised and Traditional Term Weighting Methods for Automatic Text Categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **31**, 721-735. [CrossRef]

120. Fu-ren Lin, Jen-Hung Yu. 2009. Visualized cognitive knowledge map integration for P2P networks. *Decision Support Systems* **46**, 774-785. [CrossRef]

121. Robert Jay Milewski, Venu Govindaraju, Anurag Bhardwaj. 2009. Automatic recognition of handwritten medical forms for search engines. *International Journal of Document Analysis and Recognition (IJDAR)* **11**, 203-218. [CrossRef]

122. P. Carpena, P. Bernaola-Galván, M. Hackenberg, A. Coronado, J. Oliver. 2009. Level statistics of words: Finding keywords in literary texts and symbolic sequences. *Physical Review E* **79**. . [CrossRef]

123. Takeshi TSUCHIYA, Hiroaki SAWANO, Marc LIHAN, Hirokazu YOSHINAGA, Keiichi KOYANAGI. 2009. A distributed information retrieval manner based on the statistic information for ubiquitous services. *Progress in Informatics* 63. [CrossRef]

124. Constantin Orăsan. 2009. Comparative Evaluation of Term-Weighting Methods for Automatic Summarization*. *Journal of Quantitative Linguistics* **16**, 67-95. [CrossRef]

125. Daniel Tunkelang. 2009. Faceted Search. *Synthesis Lectures on Information Concepts, Retrieval, and Services* **1**, 1-80. [CrossRef]

126. Peter Willett. 2009. Similarity methods in chemoinformatics. *Annual Review of Information Science and Technology* **43**:10.1002/aris.144.v43:1, 1-117. [CrossRef]

127. Satoru Takahashi, Hiroshi Takahashi, Kazuhiko Tsuda. 2009. Analysis of the effect of Headline News in financial market through text categorisation. *International Journal of Computer Applications in Technology* **35**, 204. [CrossRef]
128. Carlos Costa, Filipe Freitas, Marco Pereira, Augusto Silva, José L. Oliveira. 2009. Indexing and retrieving DICOM data in disperse and unstructured archives. *International Journal of Computer Assisted Radiology and Surgery* **4**, 71-77. [CrossRef]
129. Takeshi TSUCHIYA, Hirokazu YOSHINAGA, Keiichi KOYANAGI. 2009. P2P Distributed Information Retrieval Method by the Independent Indices Management for the Ubiquitous Services Environment. *Journal of Japan Society for Fuzzy Theory and Intelligent Informatics* **21**, 129-142. [CrossRef]
130. Xiaoyuan Su, Taghi M. Khoshgoftaar. 2009. A Survey of Collaborative Filtering Techniques. *Advances in Artificial Intelligence* **2009**, 1-19. [CrossRef]
131. Joris Vertommen, Frizo Janssens, Bart De Moor, Joost R. Duflou. 2008. Multiple-vector user profiles in support of knowledge sharing. *Information Sciences* **178**, 3333-3346. [CrossRef]
132. W.S. Wong, R.W.P. Luk, H.V. Leong, K.S. Ho, D.L. Lee. 2008. Re-examining the effects of adding relevance information in a relevance feedback environment. *Information Processing & Management* **44**, 1086-1116. [CrossRef]
133. Kristof Coussement, Dirk Van den Poel. 2008. Integrating the voice of customers through call center emails into a decision support system for churn prediction. *Information & Management* **45**, 164-174. [CrossRef]
134. Steve Whittaker, Simon Tucker, Kumutha Swampillai, Rachel Laban. 2008. Design and evaluation of systems to support interaction capture and retrieval. *Personal and Ubiquitous Computing* **12**, 197-221. [CrossRef]
135. Paul Thompson. 2008. Looking back: On relevance, probabilistic indexing and information retrieval. *Information Processing & Management* **44**, 963-970. [CrossRef]
136. Stephen Robertson, John Tait. 2008. Karen Spärck Jones. *Journal of the American Society for Information Science and Technology* **59**:10.1002/asi.v59:5, 852-854. [CrossRef]
137. Wa'el Musa Hadi, Fadi Thabtah, Salahideen Mousa, Samer Al Hawari, Ghassan Kanaan, Jafar Ababnih. 2008. A Comprehensive Comparative Study Using Vector Space Model with K-Nearest Neighbor on Text Categorization Data. *Asian Journal of Information Management* **2**, 14-22. [CrossRef]
138. ChengXiang Zhai. 2008. Statistical Language Models for Information Retrieval. *Synthesis Lectures on Human Language Technologies* **1**, 1-141. [CrossRef]
139. C. Lioma, I. Ounis. 2008. A syntactically-based query reformulation technique for information retrieval. *Information Processing & Management* **44**, 143-162. [CrossRef]
140. Dietmar Wolfram, Jin Zhang. 2008. The influence of indexing practices and weighting algorithms on document spaces. *Journal of the American Society for Information Science and Technology* **59**:10.1002/asi.v59:1, 3-11. [CrossRef]
141. Stephen Robertson, John Tait. 2007. In Memoriam. *Information Processing & Management* **43**, 1441-1444. [CrossRef]
142. John I. Tait. 2007. Karen Spärck Jones. *Computational Linguistics* **33**, 289-291. [CrossRef]
143. Kin Leong Ho, Paul Newman. 2007. Detecting Loop Closure with Scene Sequences. *International Journal of Computer Vision* **74**, 261-286. [CrossRef]
144. Jonathan M. Levitt. 2007. Developing and applying a metric for estimating trends in databases of articles. *Collnet Journal of Scientometrics and Information Management* **1**, 17-21. [CrossRef]

145. Yorick Wilks. 2007. Karen Sparck Jones (1935-2007). *IEEE Intelligent Systems* **22**, 8-9. [CrossRef]
146. Howard D. White. 2007. Combining bibliometrics, information retrieval, and relevance theory, Part 1: First examples of a synthesis. *Journal of the American Society for Information Science and Technology* **58**:10.1002/asi.v58:4, 536-559. [CrossRef]
147. Haizhou Li, Bin Ma, Chin-Hui Lee. 2007. A Vector Space Modeling Approach to Spoken Language Identification. *IEEE Transactions on Audio, Speech and Language Processing* **15**, 271-284. [CrossRef]
148. Marco Aiello, Andrea Pegoretti. 2006. TEXTUAL ARTICLE CLUSTERING IN NEWSPAPER PAGES. *Applied Artificial Intelligence* **20**, 767-796. [CrossRef]
149. Fu-ren Lin, Kuen-jin Huang, Nian-shing Chen. 2006. Integrating information retrieval and data mining to discover project team coordination patterns. *Decision Support Systems* **42**, 745-758. [CrossRef]
150. Giyeong Kim. 2006. Relationship between index term specificity and relevance judgment. *Information Processing & Management* **42**, 1218-1229. [CrossRef]
151. Ronan Cummins, Colm O'Riordan. 2006. Evolving local and global weighting schemes in information retrieval. *Information Retrieval* **9**, 311-330. [CrossRef]
152. Alfred Kobsa, Josef Fink. 2006. An LDAP-based User Modeling Server and its Evaluation. *User Modeling and User-Adapted Interaction* **16**, 129-169. [CrossRef]
153. Patrice Buche, Juliette Dibie-Barthélemy, Ollivier Haemmerlé, Gaëlle Hignette. 2006. Fuzzy semantic tagging and flexible querying of XML documents extracted from the Web. *Journal of Intelligent Information Systems* **26**, 25-40. [CrossRef]
154. Olga Vechtomova, Murat Karamuftuoglu. 2006. Elicitation and use of relevance feedback information. *Information Processing & Management* **42**, 191-206. [CrossRef]
155. Birger Hjørland, Karsten Nissen Pedersen. 2005. A substantive theory of classification for information retrieval. *Journal of Documentation* **61**:5, 582-597. [Abstract] [Full Text] [PDF]
156. Lin-Chih Chen, Cheng-Jye Luh. 2005. Web page prediction from metasearch results. *Internet Research* **15**:4, 421-446. [Abstract] [Full Text] [PDF]
157. Bo-Yeong Kang, Sang-Jo Lee. 2005. Document indexing: a concept-based approach to term weight estimation. *Information Processing & Management* **41**, 1065-1080. [CrossRef]
158. Lin-Chih Chen, Cheng-Jye Luh, Chichang Jou. 2005. Generating page clippings from web search results using a dynamically terminated genetic algorithm. *Information Systems* **30**, 299-316. [CrossRef]
159. Ricardo D. Fierro, Eric P. Jiang. 2005. Lanczos and the Riemannian SVD in information retrieval applications. *Numerical Linear Algebra with Applications* **12**:10.1002/nla.v12:4, 355-372. [CrossRef]
160. Vassilis Plachouras, Iadh Ounis. 2005. Dempster-Shafer Theory for a Query-Biased Combination of Evidence on the Web. *Information Retrieval* **8**, 197-218. [CrossRef]
161. Gloria Bordogna, Gabriella Pasi. 2005. Personalised Indexing and Retrieval of Heterogeneous Structured Documents. *Information Retrieval* **8**, 301-318. [CrossRef]
162. Karen Spärck Jones. 2005. Some Points in a Time. *Computational Linguistics* **31**, 1-14. [CrossRef]
163. Jin Zhang, Tien N. Nguyen. 2005. A New Term Significance Weighting Approach. *Journal of Intelligent Information Systems* **24**, 61-85. [CrossRef]
164. Mark J. Huiskes, Eric J. PauwelsIndexing, Learning and Content-Based Retrieval for Special Purpose Image Databases 203-258. [CrossRef]
165. Filippo Menczer. 2004. Lexical and semantic clustering by Web links. *Journal of the American Society for Information Science and Technology* **55**:10.1002/asi.v55:14, 1261-1269. [CrossRef]

166. NRIPENDRA L. SHRESTHA, YOUHEI KAWAGUCHI, TAKENAO OHKAWA. 2004. SUMOMO: A PROTEIN SURFACE MOTIF MINING MODULE. *International Journal of Computational Intelligence and Applications* **04**, 431-449. [CrossRef]

167. Stephen Robertson. 2004. Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation* **60**:5, 503-520. [Abstract] [Full Text] [PDF]

168. Qing Ma, Kyoko Kanzaki, Yujie Zhang, Masaki Murata, Hitoshi Isahara. 2004. Self-organizing semantic maps and its application to word alignment in Japanese–Chinese parallel corpora. *Neural Networks* **17**, 1241-1253. [CrossRef]

169. 2004. An Experimental Study on Multi-Document Summarization for Question Answering. *Journal of the Korean Society for information Management* **21**, 289-303. [CrossRef]

170. Padmini Srinivasan. 2004. Text mining: Generating hypotheses from MEDLINE. *Journal of the American Society for Information Science and Technology* **55**:10.1002/asi.v55:5, 396-413. [CrossRef]

171. J.A. Lay, L. Guan. 2004. Retrieval For Color Artistry Concepts. *IEEE Transactions on Image Processing* **13**, 326-339. [CrossRef]

172. R.M. Losee, L. Church. 2004. Information retrieval with distributed databases: analytic models of performance. *IEEE Transactions on Parallel and Distributed Systems* **15**, 18-27. [CrossRef]

173. Manabu Matsuyama, Satoshi Shintani, Takayuki Ito, Yusuke Hiraoka, Satoshi Watanabe. 2004. A Keyword Extraction Method for generating a User Profile in the Paper Collection and Sharing System MiDoc. *IEEJ Transactions on Electronics, Information and Systems* **124**, 2489-2494. [CrossRef]

174. Susan T. Dumais. 2004. Latent semantic analysis. *Annual Review of Information Science and Technology* **38**:10.1002/aris.v38:1, 188-230. [CrossRef]

175. 2003. A Study on Query Refinement by Online Relevance Feedback in an Information Filtering System. *Journal of the Korean Society for information Management* **20**, 23-48. [CrossRef]

176. 2003. A Study on the Pivoted Inverse Document Frequency Weighting Method. *Journal of the Korean Society for information Management* **20**, 233-248. [CrossRef]

177. Hong-Kwang J. Kuo, Olivier Siohan, Joseph P. Olive. 2003. Advances in natural language call routing. *Bell Labs Technical Journal* **7**, 155-170. [CrossRef]

178. El-Sayed Atlam, M. Fuketa, K. Morita, Jun-ichi Aoe. 2003. Documents similarity measurement using field association terms. *Information Processing & Management* **39**, 809-824. [CrossRef]

179. J Morato, J Llorens, G Genova, J.A Moreiro. 2003. Experiments in discourse analysis impact on information classification and retrieval algorithms. *Information Processing & Management* **39**, 825-851. [CrossRef]

180. Zhongzhi Shi, Qing He, Ziyan Jia, Jiayou Li. 2003. Intelligence Chinese Document Semantic Indexing System. *International Journal of Information Technology & Decision Making* **02**, 407-424. [CrossRef]

181. Jean-Pierre Koenig, Gail Mauner, Breton Bienvenue. 2003. Arguments for adjuncts. *Cognition* **89**, 67-103. [CrossRef]

182. Yoshikiyo Kato, Takahiro Shirakawa, Kohei Taketa, Koichi Hori. 2003. An approach to discovering risks in development process of large and complex systems. *New Generation Computing* **21**, 163-176. [CrossRef]

183. Linnea Marshall. 2003. Specific and Generic Subject Headings: Increasing Subject Access to Library Materials. *Cataloging & Classification Quarterly* **36**, 59-87. [CrossRef]

184. Karen Sparck Jones. 2003. 2002 ASIST Award of Merit: Karen Sparck Jones. *Bulletin of the American Society for Information Science and Technology* **29**, 12-14. [CrossRef]
185. Akiko Aizawa. 2003. An information-theoretic perspective of tf–idf measures. *Information Processing & Management* **39**, 45-65. [CrossRef]
186. Naohiro Matsumura, Yukio Ohsawa, Mitsuru Ishizuka. 2003. Profiling Participants in Online-Community Based on Influence Diffusion Model. *Transactions of the Japanese Society for Artificial Intelligence* **18**, 165-172. [CrossRef]
187. Ian Ruthven, Mounia Lalmas, Keith van Rijsbergen. 2002. Combining and selecting characteristics of information use. *Journal of the American Society for Information Science and Technology* **53**:10.1002/asi.v53:5, 378-396. [CrossRef]
188. R. Krishnapuram, A. Joshi, O. Nasraoui, L. Yi. 2001. Low-complexity fuzzy relational clustering algorithms for Web mining. *IEEE Transactions on Fuzzy Systems* **9**, 595-607. [CrossRef]
189. Mikio Yamamoto, Kenneth W. Church. 2001. Using Suffix Arrays to Compute Term Frequency and Document Frequency for All Substrings in a Corpus. *Computational Linguistics* **27**, 1-30. [CrossRef]
190. Robert M. Losee. 2001. Term dependence: A basis for Luhn and Zipf models. *Journal of the American Society for Information Science and Technology* **52**:10.1002/asi.v52:12, 1019-1025. [CrossRef]
191. K. Sparck Jones, S. Walker, S.E. Robertson. 2000. A probabilistic model of information retrieval: development and comparative experiments. *Information Processing & Management* **36**, 779-808. [CrossRef]
192. Jorng-Tzong Horng, Ching-Chang Yeh. 2000. Applying genetic algorithms to query optimization in document retrieval. *Information Processing & Management* **36**, 737-759. [CrossRef]
193. Vaibhava Goel, William J Byrne. 2000. Minimum Bayes-risk automatic speech recognition. *Computer Speech & Language* **14**, 115-135. [CrossRef]
194. Robert M. Losee, Lee Anne H. Paris. 1999. Measuring search-engine quality and query difficulty: Ranking with target and freestyle. *Journal of the American Society for Information Science* **50**:10.1002/(SICI)1097-4571(1999)50:10<>1.0.CO;2-I, 882-889. [CrossRef]
195. Michael W. Berry, Zlatko Drmac, Elizabeth R. Jessup. 1999. Matrices, Vector Spaces, and Information Retrieval. *SIAM Review* **41**, 335-362. [CrossRef]
196. Myoung-Cheol Kim, Key-Sun Choi. 1999. A comparison of collocation-based similarity measures in query expansion. *Information Processing & Management* **35**, 19-30. [CrossRef]
197. Alexander M. Robertson, Peter Willett. 1998. Applications of n-grams in textual information systems. *Journal of Documentation* **54**:1, 48-67. [Abstract] [PDF]
198. Lee Anne H. Paris, Helen R. Tibbo. 1998. Freestyle vs. Boolean: A comparison of partial and exact match retrieval systems. *Information Processing & Management* **34**, 175-190. [CrossRef]
199. Massimo Melucci. 1998. Passage retrieval: A probabilistic technique. *Information Processing & Management* **34**, 43-68. [CrossRef]
200. Hyun-Kyu Kang, Key-Sun Choi. 1997. Two-level document ranking using mutual information in natural language information retrieval. *Information Processing & Management* **33**, 289-306. [CrossRef]
201. ALEXANDER M. ROBERTSON, PETER WILLETT. 1996. AN UPPERBOUND TO THE PERFORMANCE OF RANKED-OUTPUT SEARCHING: OPTIMAL WEIGHTING OF QUERY TERMS USING A GENETIC ALGORITHM. *Journal of Documentation* **52**:4, 405-420. [Abstract] [PDF]

202. Scott Henninger. 1995. Information access tools for software reuse. *Journal of Systems and Software* **30**, 231-247. [CrossRef]
203. Robert M. Losee. 1995. Determining information retrieval and filtering performance without experimentation. *Information Processing & Management* **31**, 555-572. [CrossRef]
204. Kenneth W. Church, William A. Gale. 1995. Poisson mixtures. *Natural Language Engineering* **1**. . [CrossRef]
205. Tomek Strzalkowski. 1995. Natural language information retrieval. *Information Processing & Management* **31**, 397-417. [CrossRef]
206. Gloria Bordogna, Gabriella Pasi. 1995. Controlling retrieval through a user-adaptive representation of documents. *International Journal of Approximate Reasoning* **12**, 317-339. [CrossRef]
207. Jacques Savoy. 1994. A learning scheme for information retrieval in hypertext. *Information Processing & Management* **30**, 515-533. [CrossRef]
208. W.John Wilbur, Leona Coffee. 1994. The effectiveness of document neighboring in search enhancement. *Information Processing & Management* **30**, 253-266. [CrossRef]
209. Andrew Jennings, Hideyuki Higuchi. 1993. A user model neural network for a personal news service. *User Modelling and User-Adapted Interaction* **3**, 1-25. [CrossRef]
210. J.Christopher Westland. 1992. Economic incentives for database normalization. *Information Processing & Management* **28**, 647-662. [CrossRef]
211. G. SALTON. 1991. Developments in Automatic Text Retrieval. *Science* **253**, 974-980. [CrossRef]
212. Susan T. Dumais. 1991. Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, & Computers* **23**, 229-236. [CrossRef]
213. E. MICHAEL KEEN. 1991. THE USE OF TERM POSITION DEVICES IN RANKED OUTPUT EXPERIMENTS. *Journal of Documentation* **47**:1, 1-22. [Abstract] [PDF]
214. S.K.M Wong, Y.Y Yao. 1991. A probabilistic inference model for information retrieval. *Information Systems* **16**, 301-321. [CrossRef]
215. Donna Harman, Wayne McCoy, Robert Toense, Gerald Candela. 1991. Prototyping a distributed information retrieval system that uses statistical ranking. *Information Processing & Management* **27**, 449-460. [CrossRef]
216. K.L. KwokQuery Learning Using an ANN with Adaptive Architecture 260-264. [CrossRef]
217. Padmini Srinivasan. 1990. A comparison of two-poisson, inverse document frequency and discrimination value models of document representation. *Information Processing & Management* **26**, 269-278. [CrossRef]
218. W.M. Shaw. 1990. Subject indexing and citation indexing— part II: An evaluation and comparison. *Information Processing & Management* **26**, 705-718. [CrossRef]
219. W.M. Shaw. 1990. Subject indexing and citation indexing—part I: Clustering structure in the cystic fibrosis document collection#. *Information Processing & Management* **26**, 693-703. [CrossRef]
220. Gerard Salton, Chris Buckley, Maria Smith. 1990. On the application of syntactic methodologies in automatic text analysis. *Information Processing & Management* **26**, 73-92. [CrossRef]
221. Karen E. Lochbaum, Lynn A. Streeter. 1989. Comparing and combining the effectiveness of latent semantic indexing and the ordinary vector space model for information retrieval. *Information Processing & Management* **25**, 665-676. [CrossRef]
222. Michael J. Nelson. 1988. Correlation of term usage and term indexing frequencies. *Information Processing & Management* **24**, 541-547. [CrossRef]

223. G. Salton. 1988. A simple blueprint for automatic Boolean query processing. *Information Processing & Management* **24**, 269-280. [CrossRef]
224. K.L. Kwok, William Kuan. 1988. Experiments with document components for indexing and retrieval. *Information Processing & Management* **24**, 405-417. [CrossRef]
225. Gerard Salton, Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management* **24**, 513-523. [CrossRef]
226. Susan T. DumaisTextual Information Retrieval 673-700. [CrossRef]
227. Howard D. White, Belver C. Griffith. 1987. Quality of indexing in online data bases. *Information Processing & Management* **23**, 211-224. [CrossRef]
228. William P. Jones. 1986. On the applied use of human memory models: the memory extender personal filing system. *International Journal of Man-Machine Studies* **25**, 191-228. [CrossRef]
229. SUZANNE BERTRAND-GASTALDY, COLIN H. DAVIDSON. 1986. IMPROVED DESIGN OF GRAPHIC DISPLAYS IN THESAURI — THROUGH TECHNOLOGY AND ERGONOMICS. *Journal of Documentation* **42**:4, 225-251. [Abstract] [PDF]
230. Ian G Hendry, Peter Willett, Frances E. Wood. 1986. INSTRUCT: a teaching package for experimental methods in information retrieval. Part II. Computational aspects. *Program* **20**:4, 382-393. [Abstract] [PDF]
231. S.E. ROBERTSON. 1986. ON RELEVANCE WEIGHT ESTIMATION AND QUERY EXPANSION. *Journal of Documentation* **42**:3, 182-188. [Abstract] [PDF]
232. W.M. Shaw. 1986. An investigation of document partitions. *Information Processing & Management* **22**, 19-28. [CrossRef]
233. G. Salton, E. A. Fox, E. Voorhees. 1985. Advanced feedback methods in information retrieval. *Journal of the American Society for Information Science* **36**:10.1002/asi.v36:3, 200-210. [CrossRef]
234. W.Bruce Croft, Thomas J Parenty. 1985. A comparison of a network structure and a database system used for document retrieval. *Information Systems* **10**, 377-390. [CrossRef]
235. Donald H. KraftAdvances in Information Retrieval: Where Is That /#*&@¢ Record? 277-318. [CrossRef]
236. G. Salton, E. Voorhees, E.A. Fox. 1984. A comparison of two methods for boolean query relevancy feedback†. *Information Processing & Management* **20**, 637-651. [CrossRef]
237. G. Salton, C. Buckley, E. A. Fox. 1983. Automatic query formulations in information retrieval. *Journal of the American Society for Information Science* **34**:10.1002/asi.v34:4, 262-280. [CrossRef]
238. ABRAHAM BOOKSTEIN. 1983. OUTLINE OF A GENERAL PROBABILISTIC RETRIEVAL MODEL. *Journal of Documentation* **39**:2, 63-72. [Abstract] [PDF]
239. Alan F. Harding, Michael F. Lynch, Peter Willett. 1983. Document retrieval using a serial bit string search. *Information Processing & Management* **19**, 1-8. [CrossRef]
240. Kazimierz Choroś, Czesław Daniłowicz. 1982. Weighted descriptors and relative indexing in a document retrieval system model. *Information Processing & Management* **18**, 207-220. [CrossRef]
241. W. Bruce Croft. 1981. Document representation in probabilistic models of information retrieval. *Journal of the American Society for Information Science* **32**:10.1002/asi.v32:6, 451-457. [CrossRef]
242. G. Salton, H. Wu, C. T. Yu. 1981. The measurement of term importance in automatic indexing. *Journal of the American Society for Information Science* **32**:10.1002/asi.v32:3, 175-186. [CrossRef]
243. HARRY WU, GERARD SALTON. 1981. THE ESTIMATION OF TERM RELEVANCE WEIGHTS USING RELEVANCE FEEDBACK. *Journal of Documentation* **37**:4, 194-214. [Abstract] [PDF]

244. V. M. Driyanskii. 1981. Retrieval models in on-line documentary information systems: An analytic review. *Cybernetics* **17**, 269-287. [CrossRef]

245. G. Salton, R.K. Waldstein. 1978. Term relevance weights in on-line information retrieval. *Information Processing & Management* **14**, 29-35. [CrossRef]

246. S.E. ROBERTSON. 1977. THEORIES AND MODELS IN INFORMATION RETRIEVAL. *Journal of Documentation* **33**:2, 126-148. [Abstract] [PDF]

247. S. E. Robertson, K. Sparck Jones. 1976. Relevance weighting of search terms. *Journal of the American Society for Information Science* **27**:10.1002/asi.v27:3, 129-146. [CrossRef]

248. Stephen P. Harter. 1975. A probabilistic approach to automatic keyword indexing. Part II. An algorithm for probabilistic indexing. *Journal of the American Society for Information Science* **26**:10.1002/asi.v26:5, 280-289. [CrossRef]

249. KAREN SPARCK JONES. 1975. A PERFORMANCE YARDSTICK FOR TEST COLLECTIONS. *Journal of Documentation* **31**:4, 266-272. [Abstract] [PDF]

250. G. Salton, C. S. Yang, C. T. Yu. 1975. A theory of term importance in automatic text analysis. *Journal of the American Society for Information Science* **26**:10.1002/asi.v26:1, 33-44. [CrossRef]

251. M. H. Heine. 1974. Design equations for retrieval systems based on the swets model. *Journal of the American Society for Information Science* **25**:10.1002/asi.v25:3, 183-198. [CrossRef]

252. KAREN SPARCK JONES. 1974. AUTOMATIC INDEXING. *Journal of Documentation* **30**:4, 393-432. [Abstract] [PDF]

253. 1974. DOCUMENTATION NOTE. *Journal of Documentation* **30**:1, 41-46. [Abstract] [PDF]

254. Karen Sparck Jones. 1973. Index term weighting. *Information Storage and Retrieval* **9**, 619-633. [CrossRef]

255. G. SALTON, C.S. YANG. 1973. ON THE SPECIFICATION OF TERM VALUES IN AUTOMATIC INDEXING. *Journal of Documentation* **29**:4, 351-372. [Abstract] [PDF]

256. Karen Sparck Jones. 1973. Letters to the Editor. *Journal of the American Society for Information Science* **24**:10.1002/asi.v24:2, 166-167. [CrossRef]

257. 1973. DOCUMENTATION NOTES. *Journal of Documentation* **29**:1, 72-84. [Abstract] [PDF]

258. B.C. BROOKES. 1972. THE SHANNON MODEL OF IR SYSTEMS. *Journal of Documentation* **28**:2, 160-162. [Abstract] [PDF]

259. Louis Massey, Wilson WongA Cognitive-Based Approach to Identify Topics in Text Using the Web as a Knowledge Source 61-78. [CrossRef]

260. Mary HolstegeWhere Are All The Bugs? . [CrossRef]

261. Yasuhiro Yamada, Kanji Kato, Sachio HirokawaText Mining for Analysis of Interviews and Questionnaires 52-68. [CrossRef]

262. Martin Hassel, Hercules DalianisPortable Text Summarization 17-32. [CrossRef]

263. Yasuhiro Yamada, Kanji Kato, Sachio HirokawaText Mining for Analysis of Interviews and Questionnaires 1390-1406. [CrossRef]

264. Gong Cheng, Yuzhong QuSearching Linked Objects with Falcons 259-278. [CrossRef]

265. Mehmet GencerThe Evolution and Specialization of IETF Standards 66-85. [CrossRef]