Clustering Information Retrieval Search Outputs

Authors: Yasemin Kural, Steve Robertson, Susan Jones

e-mail: de535, ser, sa386 @ soi.city.ac.uk
School of Informatics, City University

Abstract

Users are known to experience difficulties in dealing with information retrieval search outputs, especially if those outputs are above a certain size. It has been argued by several researchers that search output clustering can help users in their interaction with IR systems in some retrieval situations, providing them with an overview of their results by exploiting the topicality information that resides in the output but has not been used at the retrieval stage. This overview might enable them to find relevant documents more easily by focusing on the most promising clusters, or to use the clusters as a starting-point for query refinement or expansion. In this paper, the results of experiments carried out to assess the viability of clustering as a search output presentation method are reported and discussed.

1. Introduction

A user's interaction with a search output is often far from optimal. Especially when the output exceeds a certain threshold, users are inclined to sample just a few documents or abandon the query altogether. With a Boolean system, the search output can be reduced by introducing additional search terms, but studies show that a great majority of users do very little or no Boolean searches [28]. Even experienced users may not be willing or able to find the appropriate terms to narrow down a search [30].

Most IR systems provide the users with a relevance-ranked list to help them find relevant documents easily, but in cases where a user experiences difficulties in expressing his information need, or has a more exploratory approach towards a search output, or when many documents have the same score, relevance ranking may not be very helpful. It has been proposed that clustering can help users in such cases, by showing them some kind of pattern existing in the document set, enabling them to overview the set quickly and make judgements on groups of documents simultaneously. Alternatively, if the coverage of the output is inappropriate for the user's need, the topicality information presented in the cluster representations may give cues for modifying the query. If clustering achieves a helpful categorisation of the documents, it may also act as a versatile query expansion aid.

This article reports experimental findings from a PhD project, in which output clustering was investigated as a user interaction tool, working on a limited number of documents retrieved by the Okapi probabilistic search engine. Can&Ozkarahan's [2] C^3M algorithm was used with small modifications for clustering the documents, and probabilistic methods were employed for document retrieval and cluster representations.

2. Choice Of Method

There are many different clustering methods which are neither mutually exclusive nor can be neatly categorised into a few groups. Hierarchical methods are the broadest family to be categorised under one group; see Everitt [9] for a detailed classification. In the past, much discussion on the choice of clustering methods has focused on computing efficiency, as clustering algorithms are notoriously complicated and processor-intensive. Nowadays this is of much less concern, and the various methods can be judged purely on their results.

No clustering method can be judged 'best' in all circumstances, and it is infeasible to comprehensively test a wide range of clustering methods before choosing one. In this project, the C3M method was chosen as it conformed to the theoretical soundness criteria [14], and provided a means of estimating the optimum number of clusters as well as identifying cluster centroids and forming clusters around the centroids. It also had a history of good performance with the Inspec database [3] which was available for the project experiments.

Unlike hierarchical clustering methods which have been generally favoured by CBR researchers [15, 33, 36], this method allowed *overlapping*; a feature that was found to be useful in the context of this project. In contrast to the general preference for hierarchical clustering methods in CBR, it is argued here that document collections do not have any intrinsic qualities that make them appropriate for hierarchic conceptualisation. On

the contrary, due to the variety of aspects a document may cover, an overlapping classification where a document can be a member of more than one group seems more appropriate. Everitt's [9] statement below might apply to document clustering as well:

> *"It is in biological applications such as the evolutionary trees that hierarchical classifications are most relevant. Hierarchical clustering procedures are, however, now used in many other fields in which hierarchical structures may not be the most appropriate. The danger of imposing a hierarchical scheme on data which is essentially non-hierarchical is clear."*

It is puzzling to observe CBR researchers' overwhelming preference for hierarchical methods, especially when we consider that they are more demanding in their assumptions than their non-hierarchical counterparts. Hierarchical methods attempt to create useful groups out of a set of items at various different levels: for a document set of 100 documents, this implies that the documents can be divided into, say, 2, 4, 7, 10, 15 and 30 groups in a meaningful way.

In fact, research work to date shows that even when a hierarchical structure is preferred, it is typically only the bottom-level clusters which give good retrieval results [8, 11, 12, 32, 27]. Even if we have a hierarchical representation, we do not seem to have much use for its upper levels. This is quite intuitive when we consider the amount of information reduction involved there.

For this project, the maximum size of the document sets to be clustered was set at 50. This figure was large enough to make it worthwhile to apply clustering, and small enough to cluster without excessive information reduction. In these circumstances we needed nothing more than a simple partitioning method, making the use of the hierarchical method even less appropriate.

3. Other Decisions Involved In Output Clustering

When implementing a clustering method, it is first necessary to decide on the type and number of variables to be used. Documents are generally represented by *terms* for clustering purposes, but it has been argued that relevance cannot be limited to topicality; and a variety of other factors such as authors, journal, obtainability, cost, and previously seen documents all affect user decisions [1, 10]. In this implementation, the document representation was limited to term occurrences as most of these other factors were quite difficult to measure and incorporate in the implementation.

As clustering results would be affected by the number of terms used, various experiments were done to find the approximate number of terms that could be expected to produce a balanced distribution of documents among clusters, and set upper and lower boundaries for the algorithm. The lower boundary was set at four, since terms occurring only in one or two documents would not be of any use at all.

Setting the upper boundary involved more thought. Obviously the more terms are used the more information is available for the clustering algorithm, but the greater the risk of obscuring the cluster structure [9, 18]. One method to reduce this risk was to use a list of about 900 stopwords in conjunction with a stemming algorithm, to weed out non-contextual terms. In addition, a long list of synonyms (about 950) was employed in order not to double-count words that could be used interchangably, or to let similarities between documents remain unexploited due to different wordings of the same expression or idea.

The total number of terms was then limited to a maximum of 70, but after applying the above methods the number of eligible candidates usually fell somewhat short of this figure. In cases where it was exceeded, the candidate terms with highest Term Selection Value [23] were chosen. TSV was a more appropriate criterion for our purposes than term weight, since it correlates with a term's ability to discriminate between document groups.

It was also necessary to decide whether to use the algorithm in weighted or binary mode. Binary mode was chosen as it was simpler to implement, and C3M was reported to perform well with it.

Another group of decisions involved producing concise representations of cluster topics for users' viewing. Such representations could in principle consist of representative *terms* and / or representative *titles*, and it was necessary to establish how to select those which would provide the best means of discrimination, and enable users to assess the clusters most easily.

After some experiments, it was decided to present a combination of three titles and up to ten terms, i.e. those with the highest TSVs. (For an example see Table 1 below.) In order to maximise discrimination, only those representative terms that did not occur in any other cluster representation were selected for display. Representative terms on their own were found to be somewhat cryptic for participants to use in evaluating clusters, so, in order to show them within a meaningful context, it was decided to choose those representative titles containing the most representative terms, as well as listing the terms themselves.

The alternative to this selection criterion was to choose representative titles based on their similarity to the cluster *seed* document, ensuring that the titles of the most typical members of the clusters would be shown, regardless of whether they conveyed a good representation of the document contents.

In order to test these two alternative approaches, ten preliminary experiments were performed. These

revealed that users found the titles to be more important than the terms in assessing clusters, and that choosing the titles according to the number of query and representative terms they included produced a viable representation. However there was no significant difference between the two title selection methods in terms of user preference, and a decision was made to use both in subsequent experiments. Users would be asked to rank the clusters in each representation separately, and, at the end of the experiment, to compare them in their perceived usefulness.

Table 1 : Example cluster representation

(Query: IR, system, evaluation, performance, compare, criteria, comparative)

….
CLUSTER 2 (includes 15 documents) RANK( )
REPRESENTATIVE DOCUMENTS
5: A critical investigation of recall and precision as measures of retrieval system performance
17: On probabilistic notions of precision as a function of recall
39: Implementation and evaluation of a relevance feedback device based on neural networks.
REPRESENTATIVE TERMS
probabilistic - document - system - query - investigation - relevant - retrieval - evaluation -
computerised information retrieval – accuracy
CLUSTER 3 (includes 9 documents)
REPRESENTATIVE DOCUMENTS
47: Comparative modeling and evaluation of CC-NUMA and COMA on hierarchical ring architectures.
18: Latency analysis of CC-NUMA and CC-COMA rings
11:Newton: Performace improvement through comparative analysis

4. Experimental Setup

The main purpose of the experiments was to discover whether clustering could be superior to relevance ranking as a search output presentation method. As relevance ranking could not co-exist with a clustering scheme, it was necessary to evaluate the performance of clustering against ranked retrieval before being able to propose it as an alternative. The hypotheses were:

> **Null hypothesis :** The precision of relevance ranking cannot be improved by using clustering as a search output presentation method.
>
> **Alternative hypothesis :** Clustering search output can improve the precision of ranked retrieval by creating recognizable "relevant clusters" that include significantly higher proportions of relevant documents than the ranked output at comparable threshold levels[1].

A total of 85 user experiments, based on users' own information needs, have been conducted to test these hypotheses. After the first 20 experiments, performance results and user feedback were evaluated to find out ways to improve the implementation. As a result of this evaluation, some modifications were made in cluster and document representations and ten experiments were conducted to compare two alternative methods for selecting representative titles as described above. Finally, 55 experiments were conducted to attain statistically meaningful results. The results from those experiments are reported in this paper.

The first set of experiments were statistically inconclusive and as the implementation and the experimental set-up were slightly modified afterwards, their results have not been consolidated with the final results.

The general flow of the experiments was as follows:

Users were asked to write down their information need (query terms) in a pre-questionnaire, and a query was run on the Okapi search engine based on that need.

The top 50 documents[2] retrieved were clustered and users were asked to rank these clusters in order of

[1] This would also mean that a clustered output contains "irrelevant clusters" that include lower proportions of relevant documents than the ranked output at similar threshold levels.

[2] It could be desirable to use 70-100 documents in the experiments, however, in order to be able to attract participants, it was

preference.

They were then shown individual documents (titles, authors, source, date and abstract) and asked to mark each document as relevant or non-relevant.

The precision values of the clusters were then compared to the precision values of the ranked lists at comparable threshold levels. (If the cluster ranked first by the user had 12 documents, it was compared with the precision value of the ranked list at the top 12 documents level. A similar comparison was made for the documents included in the first- and second-ranked clusters.)

As output clustering represented an overhead both for the system (time and computing resources needed to perform the clustering) and the user (time needed to evaluate the cluster representations), it was important to assess whether any benefits brought about by clustering outweighed the accompanying overhead.

5. Analysis Of The Results

Almost all the participants were City University postgraduate students, most of whom had online searching experience. After each experiment they were asked to fill in a questionnaire which included questions about various aspects of the experiment. About half the users found the clustered representation useful, and in a further 25% of cases they were mildly positive about its usefulness. In 20% of the cases they thought it was not useful and in 5% they were uncertain.

When asked about their preference between the two cluster representations, in 53% of the cases, users preferred the representation where titles were chosen based on the occurrence of query and representative terms (Rep-E). The one where titles were chosen with respect to their similarity to the cluster seed (Rep-D) was preferred in 35% of the cases; in the remainder no firm preference was expressed. However, the less preferred representation (Rep-D) achieved better results in terms of precision, and thus results from Rep-D have been used in precision comparisons with the relevance ranked lists.

No significant difference was found between the ranked lists and the clustered representation when the precision values for the top cluster and the top two clusters were compared. However, although the number of cases where each method yielded higher precision was almost the same (Table 2), at the top cluster level there was a 10% difference between the average precision values in favour of the ranked lists (

Table 3). The Wilcoxon test gave 2-tailed probability of 10% at this level.

Table 2 : Comparison of ranked list with the clustered representation - number of cases

Number of cases where highest precision is provided by:	Top cluster level		Top 2 clusters level	
	number	ratio	Number	ratio
Relevance ranked list	22	40%	20	36%
Clustered representation(Rep-D)	21	38%	23	42%
Equal precision values	12	22%	12	22%
Total	55		55	

Table 3 : Comparison of ranked list with the clustered representation - average precision

Average precision:	Top cluster level	Top 2 clusters level
Relevance ranked list	55%	49%
Clustered representation(Rep-D)	50%	47%

It was also of interest to assess the extent to which users succeeded in identifying the highest precision clusters. In fact this occurred in only 16 of the 55 experiments (29%). In 19 experiments they ranked the cluster with the second best precision value first, and in 13 of those cases ranked the best cluster second. Only in 5 cases did users rank the 4th or 5th best cluster as the first (Table 4).

Table 4 : Users' ranking of highest precision clusters

crucial not to make the experiments too tiring ot time-consuming for the prospective participants, and for this reason the output size was limited to 50 documents.

User ranks the highest precision cluster:	Number	Ratio
1st	16	29%
2nd	19	35%
3rd	13	24%
4th – 5th	5	9%
Several clusters are ranked 1st by the user, or have equal precision	2	3%

Interestingly, users seemed to be better at identifying *ir*relevant clusters than relevant ones: in 19 out of the 51 cases where all clusters were ranked (37%), they ranked the lowest precision cluster last[3].When the precision values for the clusters marked last were compared to the precision values for the ranked lists at the same threshold level, it was found that these clusters often had lower precision than the ranked list; i.e. assuming that the cluster marked last by the user had *n* documents, it contained fewer relevant documents than last *n* members of the ranked list. The comparisons were made both at the last cluster and last two clusters levels. At the last cluster level, there was a significant difference between the average precision values (Wilcoxon test gave 2-Tailed P of 3.4%), but the differences at the last 2 clusters level was insignificant (Table 5).

Table 5 : Comparative precision values for cluster(s) marked last vs. ranked list

	Ranked list	Clusters	Equal
Average precision:			
at the last cluster threshold	40%	33%	
at the last 2 clusters threshold	37%	36%	
Number of cases where precision is higher:			
at the last cluster threshold	24(47%)	15(29%)	12(24%)
at the last 2 clusters threshold	25(49%)	17(33%)	9(18%)
n=51			

This tendency of the users to identify the lowest precision cluster raised the possibility of using clustering as a *rejection* rather than a selection aid, since the experiment results implied that there would be fewer relevant documents in the last cluster than at the bottom of the ranked list. If we excluded the members of this cluster from the ranked list, it was possible that other low-ranked but relevant documents could rise to higher ranks and improve overall precision. To test this idea, for the 51 cases where users actually identified a last cluster, its members were excluded from the ranked lists, which were then re-evaluated for precision. The results were analysed both at the top cluster and top two clusters level.

In practice, excluding these documents did not always increase the precision of the ranked lists. Even though some irrelevant documents (as well as some relevant ones) were removed from higher ranks they were not always replaced by other relevant documents. At the top cluster level, the precision of the ranked list was higher in 16 cases and lower in 13 cases after this process. At the top two clusters level, the precision was higher in 18 cases and lower in 14 cases (Table 6). However the comparison to ranked lists before the exclusion of the last cluster revealed a good level of improvement at the top cluster level (Table 7).

Table 6 : Precision of the ranked list after exclusion of documents from lowest ranked cluster

Effect of exclusion on precision of ranked list	Top cluster level	Top 2 clusters level
Higher	16	18
Lower	13	14
Unchanged	22	19

Table 7 : Precision of ranked lists before/after exclusion of lowest ranked cluster from the ranked list

[3] In four of the 55 cases users marked only best clusters or marked all clusters as first or second.

Number of cases where precision is higher with:	Top cluster	Top 2 clusters
Ranked list after exclusion of last cluster	24	19
Ranked lists before exclusion of last cluster	20	18
Equal	7	14
n=51		

When the ranked lists' precision values (after excluding the documents from the lowest ranked cluster) were compared with the performance of clusters marked first and second by the users, it was found that ranked lists performed as well as or better than the clusters in 76% of the cases at top cluster level (Table 8). But, at the top two clusters level, clustered representation seemed to perform slightly better than the ranked lists.

Table 8 : Precision of clusters vs ranked lists after exclusion of lowest ranked cluster from the ranked list

Number of cases where precision is higher with:	Top cluster	Top 2 clusters
Ranked list after exclusion of last cluster	24(47%)	19(37%)
Clusters	12(24%)	23(45%)
Equal	15(29%)	9(18%)

This represented some improvement in the ranked lists over their previous performance (see Table 2) against the clustered output at the top cluster level. After the exclusion of clusters marked last by the users, the precision values of the ranked lists were significantly higher than those of the clusters marked first by the users (Wilcoxon test gave a 2-tailed P value of 2.4%).

6. Significance Of Best Precision Clusters

In some recent studies, researchers have concluded that users are capable of identifying the best (i.e. highest precision) clusters, and have based some of their performance comparisons on this assumption[13]. As seen from Table 4, our experimental findings do *not* support this assumption, and reveal that users cannot be relied on to identify the best clusters. This raises questions about the validity of using best precision clusters in assessing clustering solutions. Moreover comparison of best precision clusters against the ranked lists has another flaw: it gives the clustered output an unfair advantage, as the following discussion indicates.

When the best precision clusters were compared to the ranked lists for each of the 55 user experiments, it was found that they clearly outperformed the ranked lists in terms of precision (Table 9).

Table 9 : Performance of best clusters vs ranked lists

Number of cases where Higher precision provided by:	Top cluster level	Top 2 clusters level	Total
Best cluster(s)	33(60%)	39(71%)	72(65%)
Ranked lists	9(16%)	8(15%)	17(15%)
Equal	13(24%)	8(15%)	21(19%)
n=55			

However, this remarkable performance has little practical significance, since even randomly- created clusters are likely to outperform ranked lists when sorted in precision order. The reason is that the cluster sizes are not large enough to have a distribution of relevant documents that converge to the average figures, and divergence from the average produces both low- and high-precision clusters. The smaller the cluster sizes, the higher is the chance of outperforming the ranked list. This is because:

- the effect of divergence is more pronounced: one extra relevant document makes a bigger difference to precision in a set of six documents than in a set of 20 documents,
- given a fixed number of documents, the more clusters there are, the more choices to select from.

To clarify this point, an experiment was performed to assess the extent to which clustering formed groups of documents with higher precision values than those could be expected under a random distribution. For each of the 55 queries, 100 random cluster distributions were created, with cluster sizes matching those originally

created. Precision values were calculated for each of the clusters from these distributions, and the highest values were averaged to generate an approximate expected precision for the best clusters. These values were then compared to the actual best precision values achieved in the experiments.

In 30 (55%) out of the 55 cases, the original best clusters were outperformed by the average best precision value expected under random distribution. In the remaining 25 cases, the original best clusters gave higher precision values. However, although the difference in terms of number of cases was in favour of the random distribution, the original best clusters had on average 3% better precision than the values expected under random distribution.

7. Discussion Of Results

The null hypotheses for the user experiments could not be rejected. Clustering did not in general provide users with more relevant documents than ranked lists at comparable thresholds. The precision values were quite similar, and the differences were statistically insignificant. In fact at the top cluster level, the average precision values for relevance-ranked lists were 10% better than those for the clustered outputs (55% versus 50%). Taking into account the computing overhead involved in the creation of clusters, and the additional user effort involved in assessing them, ranked lists appear to be preferable to clustered representation in terms of performance.

However, the experiments did lead to an interesting and potentially useful finding:

- there were more relevant documents at the bottom of the ranked lists compared to the clusters ranked last by the user (2-Tailed P=3.4%).
- ranked lists had significantly higher precision as compared to the clusters ranked first by the user when the clusters ranked last were excluded from the ranking (2-Tailed P=2.4%).

Thus, clustering seems more efficient as a rejection aid than a selection aid. As such, it can be used as a way of reducing the output size to be reviewed, or possibly as a source of terms to be weighted *negatively* in query expansion.

It is likely that the main reason why clustering does not fullfil our expectations is the degree of information reduction involved. An abstract already represents a level of reduction which may prevent users making valid assessments of what the complete document will tell them; a cluster representation is a twofold reduction, in the sense that it is a representation of representations (i.e. abstracts), and that it involves considerably more information reduction. We might also suggest that a successful clustering algorithm must group documents in a way that to some extent matches the user's intuitions about meaningful document categories or *aspects*.

8. Limitations Of The User Experiments

The user experiments were beneficial for assessing clustering in the context of real information needs. However they were based on the assumption that users would be interested in only one or two aspects of the output set, and judge as relevant only documents representing those aspects. Yet a well-clustered output could legitimately have relevant documents evenly distributed among its clusters, provided each cluster represented a distinct aspect or facet of the retrieved documents. To judge this function of a clustering algorithm, it would be necessary to make assessments based on topicality rather than relevance, and to investigate whether clusters were capable of grouping together documents covering distinct aspects of an output set.

However, it is difficult to conduct such experiments 'live', as they demand more time and effort from the users, and involve less realistic experimental settings. Some experiments were conducted using TREC data to assess the ability of the algorithm in this respect without user involvement, but they are not reported here.

Users' answers to the post-questionnaire also revealed some experimental limitations. In about 1/3 of the experiments, users' perceptions of information needs had changed; they stated that they could make different assessments if they had gone through the experiment again. This 'learning effect' is an inherent limitation of every IR experiment as independence of judgement cannot be assumed even for a single set of judgements over a set of documents, and unfortunately no evaluation method that takes this into account has been devised yet [29]. It is possible to minimise this effect, for example by assigning the same query to more than one user, where each user receives one type of output. But this makes the experimental setting less realistic, and also introduces complications, such as the need to account for differences between the participants.

Additionally, in about 60% of the cases, users considered one or more of the other factors (author, journal/availability, and publication year of the documents) in making their relevance judgements. These factors were not accommodated in the experimental design. While we may assume that topicality is the most important factor in users' relevance judgements, it is necessary to give some attention to these other factors, which clearly have a place in user's decision framework. However, this requires further studies into how these criteria should be weighted and accommodated in document representations.

Another weakness of the user experiments was that the evaluation methods had to be different for cluster

representations and ranked lists. Users were asked to rank the cluster representations whereas they were asked to make a "yes/no" decision for the list of documents. Especially when users are generous about marking documents as relevant, the two modes of evaluation may not be altogether comparable and the relationship between the cluster rankings and document judgements becomes blurred. This situation was most evident in cases where users marked almost all documents as relevant while giving different ranks to each of the clusters. Two such experiments, where users marked more than 40 documents as relevant, were excluded from analysis as they had little comparative value. A different method such as asking users to rank the documents instead of making a yes/no decision was not considered to be feasible as it would be too demanding of users' time and effort.

Finally, it is worth mentioning that clustering needs to be customised for different databases. As mentioned above, some experiments were done using the TREC database to investigate the relationship between clusters and topic facets. However the implementation parameters used for Inspec records did not provide balanced clustering solutions with TREC documents. These documents were more varied in size, and often considerably larger than the Inspec documents; they were also less focused, so that substantial changes were needed in order to achieve acceptable clustering. The fact that an algorithm that provides reasonable results with one database is less successful with another seems to be a good indicator of the complexities involved in developing and exploiting clustering techniques.

9. References

1. Barry C. The identification of user criteria and document characteristics: beyond the topical approach to information retrieval. Ph.D. dissertation. Syracuse, NY: Syracuse University, School of Information Studies, 1993.

2. Can F, Ozkarahan E. Concepts and effectiveness of the cover-coefficient based clustering method for text databases, ACM Transactions on Database Systems, 1990, Vol.15 N.4, 483-517.

3. Can F, Ozkarahan E. Two partitioning type clustering algorithm, JASIS, 1984, 268-276.

4. Croft WB. A model of cluster searching based on classification, Information Systems, 1980, Vol.5, 189-195.

5. Cuadra CA, Katter RV. Experimental studies of relevance judgements final report. Volume 1: Project summary. Santa Monica, CA: System Development Corp., 1967.

6. Cutting DR, Karger DR, Pedersen JO. Constant interaction time scatter-gather browsing of very large document collections, Proceedings of the 16th International ACM-SIGIR Conference on Research and Development in Information Retrieval, 126-134, 1993.

7. Cutting DR, Karger DR, Pedersen JO, et al. Scatter-gather: a cluster based approach to browsing large doc collections, Proceedings of the 15th International ACM-SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, 1992, 318-329.

8. El-Hamdouchi A, Willett P. Comparison of hierarchic agglomerative clustering methods for document retrieval, Computer Journal, 1989, Vol.32 N.3, 220-7.

9. Everitt BS. Cluster analysis, Hodder&Stoughton, 3rd ed.,1993.

10. Froehlich TJ. Relevance reconsidered - Towards an agenda for the 21st century: introduction to special topic issue on relevance research, Journal of the American Society for Information Science, 1994, Vol.45, No.3, 124-134.

11. Griffiths A, Robinson L, Willett P. Hierarchic agglomerative clustering methods for automatic document classification, Journal of Documentation, 1984, Vol.40 N.3, 175-205.

12. Griffiths A; Luckhurst; Willett P. Using inter-document similarity information in document retrieval systems, JASIS, 1986, Vol.37, 3-11.

13. Hearst M, Pedersen I. Xerox TREC-4 site report. In: Harman, DK, ed., The Fourth Text REtrieval Conference, Gaithersburg, 1996, MD: NIST.

14. Jardine N, Van Rijsbergen CJ. The use of hierarchical clustering in information retrieval, Information Storage & Retrieval, 1971, Vol.7, 217-240.

15. Jardine N, Sibson R. Mathematical taxonomy, Wiley, London&NY, 1971.

16. Jones S, Walker S, Gatford M, Do T. Peeling the onion: Okapi system architecture and software design issues, Journal of Documentation, 1997, Vol.53 N.1, 58-68.

17. Kirriemuir E, Willett P. Identification of duplicate and near-duplicate full-text records, Program, 1995, Vol.29 N.3, 241-256.

18. Lewis D. An evaluation of phrasal and clustered representations on a text categorisation task, Proceedings of the 15th International ACM-SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, 1992, 37-50.

19. Miller GA. The magical number seven, plus or minus two. Some limits on our capacity for processing information, Psychological Review, 1956, V.63, 81-97.

20. Mizzaro S. Relevance: the whole history, JASIS, 1997, V.48, N.9, 810-832.

21. Porter MF. An algorithm for suffix stripping, Program, Vol.14, No.3, 130-7, July 1980.

22. Robertson SE, Beaulieu M. Research and evaluation in information retrieval, Journal of Documentation, 1997, Vol.53 N.1, 51-57.

23. Robertson, SE. On term selection for query expansion, Journal of Documentation, 1990, Vol.46 N.4, 359-364.

24. Robertson, SE. The methodology of information retrieval experiment. In: K. Sparck Jones, ed., Information Retrieval Experiment., Butterworth&Co., London, 1981, 9-31.

25. Robertson, SE. Ranking in principle, Journal of Documentation, 1978, Vol.34, No.2, 93-100.

26. Robertson SE, Sparck Jones K. Relevance weighting of search terms, Journal of the ASIS, 1976, Vol.27, 129-46.

27. Shaw WM. An investigation of document structures, IP&M, 1990, Vol.26, 339-348.

28. Siegfried S, Bates MJ, Wilde DN. A profile of end-user searching behaviour by humanities scholars: the Getty online searching project report #2, JASIS, 1993, V.44 N.5, 273-291.

29. Sparck Jones K. Information Retrieval Experiment, London, Butterworths, 1981.

30. Su LT. On the relevance of recall and precision in user evaluation, JASIS, 1994, V.45 N.3, 207-217.

31. Tague JM . The pragmatics of information retrieval experimentation, in: Information Retrieval Experiment, ed. K. Sparck Jones, Butterworth&co., London, 1981, p. 59-102.

32. Van Rijsbergen CJ, Sparck Jones K. A test for the separation of relevant and non-rel docs in experimental test collections, Journal of Documentation, 1973, V.29, 251-257.

33. Van Rijsbergen CJ. Information Retrieval, Butterworths, London, 1979.

34. Wiberly SE, Daugherty RA, Danowski JA. User persistence in scanning postings of a computer driven information system, LCS. Library & Information Science Research (Bay38), 1990, V.12, 341-353.

35. Wiberly SE, Daugherty RA. User's persistence in scanning list of references, College & Research Libraries, 1988, V.49 N.2, 149-156.

36. Willett P. Recent trends in hierarchic document clustering: a critical review, IP&M, 1988, Vol.24 N.5, 577-597.