Inf Retrieval (2006) 9: 33–53
DOI 10.1007/s10791-005-5720-6

Hierarchical clustering of a Finnish newspaper article collection with graded relevance assessments

Tuomo Korenius · Jorma Laurikkala · Martti Juhola · Kalervo Järvelin

Received: January 26, 2004 / Revised: December 10, 2004 / Accepted: December 13, 2004

Abstract Search facilitated with agglomerative hierarchical clustering methods was studied in a collection of Finnish newspaper articles ($N = 53{,}893$). To allow quick experiments, clustering was applied to a sample ($N = 5{,}000$) that was reduced with principal components analysis. The dendrograms were heuristically cut to find an optimal partition, whose clusters were compared with each of the 30 queries to retrieve the best-matching cluster. The four-level relevance assessment was collapsed into a binary one by (A) considering all the relevant and (B) only the highly relevant documents relevant, respectively. Single linkage (SL) was the worst method. It created many tiny clusters, and, consequently, searches enabled with it had high precision and low recall. The complete linkage (CL), average linkage (AL), and Ward's methods (WM) returned reasonably-sized clusters typically of 18–32 documents. Their recall (A: 27–52%, B: 50–82%) and precision (A: 83–90%, B: 18–21%) was higher than and comparable to those of the SL clusters, respectively. The AL and WM clustering had 1–8% better effectiveness than nearest neighbor searching (NN), and SL and CL were 1–9% less efficient that NN. However, the differences were statistically insignificant. When evaluated with the liberal assessment A, the results suggest that the AL and WM clustering offer better retrieval ability than NN. Assessment B renders the AL and WM clustering better than NN, when recall is considered more important than precision. The results imply that collections in the highly inflectional and agglutinative languages, such as Finnish, may be clustered as the collections in English, provided that documents are appropriately preprocessed.

Keywords Hierarchical clustering · Graded relevance · Finnish language · Principal components analysis

T. KORENIUS · J. LAURIKKALA · M. JUHOLA
Department of Computer Sciences, University of Tampere, FIN-33014 University of Tampere, Finland
e-mail: {tuomo.korenius, jorma.laurikkala, martti.juhola}@uta.fi

K. JÄRVELIN
Center for Advanced Studies, University of Tampere, FIN-33014 University of Tampere, Finland
e-mail: kalervo.jarvelin@uta.fi

1. Introduction

Cluster analysis (Jain and Dubes 1988, Kaufman and Rousseeuw 1990, Sharma 1996, Everitt et al. 2001) is a multivariate statistical method that groups a set of objects, such as a collection of documents, so that each object typically belongs to one group (or cluster) only and the union of the groups contains all the objects. Grouping should be such that objects within a particular group have higher associations between each other than between the objects of the other groups. Cluster analysis is a well-established method that has numerous applications in the areas of archaeology, astronomy, biology, medicine, market research, and psychiatry, among others (Willett 1988, Everitt et al. 2001). Cluster analysis is also considered as one of the unsupervised machine learning methods, which do not use class information in learning. Supervised methods, such as the decision rule and tree generators (Mitchell 1997), build classification models from class labeled data. Since the composition and possibly the number of groups are in clustering unknown *a priori*, the supervised statistical and machine learning methods are fundamentally different from the cluster analysis technique. Both unsupervised and supervised methods have extensively been applied in information retrieval (Belew 2000, Sebastiani 2002). Most of the cluster analysis research in the area of information retrieval has considered the clustering of documents (Willett 1988, Baeza-Yates and Ribeiro-Neto 1999), the clustering of terms Salton (1983), and query expansion (Baeza-Yates and Ribeiro-Neto 1999).

The aim of the present study was to improve search effectiveness by clustering documents with the hierarchical clustering methods. Because of the quadratic time and memory requirements, the hierarchical clustering methods are ill-suited to highly dynamic tasks, which are common in modern information retrieval. However, the applications we are interested in are quite stationary and major real-time updates are not required. Therefore, accurate but computationally intensive methods are feasible. We selected the standard agglomerative hierarchical methods, because they have earlier been used in information retrieval and they are known to produce solutions of high quality (Willett 1988, El-Hamdouchi and Willett 1989). Furthermore, since clustering was applied and evaluated in a novel context, we felt that these well-established methods were the safest choice.

We clustered a collection of Finnish newspaper articles (Kekäläinen 1999) and evaluated the search results with a graded relevance scale (Sormunen 2000, Kekäläinen and Järvelin 2002) which judges documents as irrelevant or marginally, fairly, or highly relevant to a query. To our knowledge, the present study is the first one where a Finnish document collection has been clustered. Finnish is very different from the major European languages, and, therefore, its characteristics need to be addressed. Moreover, it seems that earlier studies have not addressed graded relevance in the evaluation of cluster-based searches. The graded assessments allow fine-grained studies and comparisons of retrieval system performance. This is especially important in modern information retrieval where an abundance of information is available, but a user is interested in only the highly relevant information, which typically is only a fraction of all the relevant information. The traditional binary relevance assessment is liberal, because it does not quantify the degree of relevance, but merely informs whether a document is relevant or irrelevant to a query. This type of approach may be infeasible, if relevant documents have clearly different levels of relevance: A small piece of relevant information is hardly as valuable to a user as a document entirely of the topic, unless one is searching for a simple fact. However, graded relevance assessment has only recently been studied in a wider scale (Sormunen 2000, 2002, Voorhees 2001, Kekäläinen and Järvelin 2002).

Since the clustering methods applied here were quite complex, the dimensions of the collection were reduced with the principal components analysis technique (Jolliffe 1986,

Sharma 1996, Rencher 2002). The hierarchical cluster structures were not used as search trees. Instead, they were cut with a heuristic rule and cluster-based searches were purely performed with the resulting partitions. It appears that this type of approach to facilitate the cluster-based search has rarely been studied earlier. Our results showed the average linkage and Ward's methods better than the single and complete linkage techniques. The searches facilitated with these two methods were able to retrieve the relevant documents better than the conventional best-match search, and the retrieved clusters were reasonably large and of good quality. These results are encouraging and show that the hierarchical cluster analysis is useful in information retrieval, provided that the environment is not highly dynamic. The results also indicate that if the features of a compound-rich and highly agglutinative language are appropriately considered, document collections in these languages may be successfully clustered.

The study is organized as follows. In Section 2, the early and current document clustering research is reviewed. Then, the hierarchical clustering methods, techniques to identify the optimal partition, principal components analysis, and the retrieval performance evaluation are described in Section 3. In Section 4, the characteristics of the Finnish language are shortly discussed and the collection and its numerical representation are presented. The results are given in Section 5 and they are discussed in Section 6, where some conclusions are also drawn and possible future work is described.

2. Document clustering

Document clustering is based on the co-occurrence information in the documents. Usually the *tf·idf* term weights are used (Willett 1988, Salton 1989), but also other co-occurrence information may facilitate document clustering. For example, citations have been used to get insights into the literature of a field, and topical themes among WWW pages have been identified on the basis of links (Rasmussen 1992, Baeza-Yates and Ribeiro-Neto 1999). Since terms may be clustered using the documents where the terms co-occur (Salton 1983, Willett 1988), document and term clustering are closely related. A recent study by Slonim and Tishby (2000) describes a double clustering procedure which actually combines these two types of clustering. Original motivations for the clustering of documents were both the increased effectiveness and efficiency of retrieval. The argument on behalf the effectiveness is based on the van Rijsbergen's cluster hypothesis which states that "closely associated documents tend to be relevant to the same requests" (van Rijsbergen 1980, p. 45). If the hypothesis is valid, the relationships between the document contents are captured by the document clusters and search is more effective, because the relevant and irrelevant clusters contain most of the relevant and irrelevant documents, respectively (Willett 1988). In addition, a search strategy that utilizes cluster structure is more efficient than the conventional exhaustive best-match search.

Most of the early clustering work in information retrieval was done with nonhierarchical methods due to the limited computing resources, but as increasingly efficient computers became available in the 1980s, the research concentrated on the hierarchical clustering methods (Rasmussen 1992). In retrospect, the experimental settings of the first clustering studies were unrealistic and their results too optimistic (Willett 1988). These studies involved mainly the Cranfield collection, which is better suited to clustering than the collections available at the time, and tiny clusters characteristic to the single linkage method were considered as satisfactory search results. Studies with other collections and hierarchical methods showed that even if clustering was less successful than assumed, it could produce result sets both of sensible size

and quality. However, the use of standard agglomerative hierarchical algorithms was found difficult in practice due to the running times of $O(N^2)$ at best and typical space requirements of $O(N^2)$, where N is the number of documents (Willett 1988, Rasmussen 1992).

Undoubtedly, the intuitive appeal of search through the tree-like cluster structure was one reason behind the past popularity of the hierarchical methods in document clustering. The search can proceed in a tree in a top-down or bottom-up manner (Willett 1988, Salton 1989). The top-down search enters the tree via its root or a node at a lower level, when top-level clusters do not differ enough. The query is then matched against the clusters (nodes) and the path of the higher similarity is selected until a stopping criterion is fulfilled. However, traversing the tree upwards from its base is a better approach, except in hierarchies built with the complete linkage algorithm (Willett 1988). Research started to move away from the cluster-based search in the early 1990s, when it was found only comparable to the simpler search methods. El-Hamdouchi and Willett (1989) showed in their careful comparison that the best-match search and the search enabled with the simple nearest neighbor clustering were usually better than four bottom-up search strategies. The popularity of the hierarchical methods declined further as Cutting et al. (1992) proposed a new paradigm which advocated the use of cluster analysis as a tool to browse large collections. Since on-line browsing requires fast responses to user's actions, hierarchical methods were too slow for new applications. In addition, along with the growth of Internet, the size of a typical collection suddenly increased from a few thousands to tens of thousands documents.

The optimization techniques (Everitt et al. 2001), such as the k-means algorithm, have been the most popular nonhierarchical clustering methods in information retrieval (Rasmussen 1992). These methods produce a flat cluster structure known as the partition by optimizing some numerical criterion (Rasmussen 1992, Everitt et al. 2001). With running times of $O(KN)$, where K is the number of clusters, they can cluster large collections reasonably fast, provided that $K \ll N$ (Rasmussen 1992). Unfortunately, these methods often assume that the user guesses the optimal number of groups and supplies starting points (seeds). Since the grouping also depends on the order in which objects are processed, the partitions may markedly differ from one run to another. Optimization techniques are not free of the problems caused by high dimensionality either. Many of them utilize cluster representatives, which may not significantly differ from each other in high dimensional data, especially when the seeds are poorly chosen (Boley 1998, Nilsson 2002).

For these reasons, hybrid clustering schemes and new algorithms have been presented. Buckshot and fractionation (Cutting et al. 1992)—the two Scatter/Gather clustering methods—are fast optimization algorithms of $O(KN)$ which utilize the more accurate, but slower average linkage algorithm to find seeds of better quality. The principal direction divisive partitioning algorithm (Boley 1998) and its non-greedy version (Nilsson 2002) are efficient methods that partition a collection in a divisive manner and produce a hierarchical tree structure which resembles the trees built by the hierarchical methods. Besides being fast, these methods have also the advantage of being deterministic, which the Scatter/Gather methods are not.

3. Methods

3.1. Hierarchical clustering

The single linkage, complete linkage, group average linkage, and Ward's clustering methods (Jain and Dubes 1988, Rasmussen 1992, Everitt et al. 2001) were selected, because they

are accurate and have often been applied in the cluster-based search studies (Willett 1988, El-Hamdouchi and Willett 1989, Rasmussen 1992). Furthermore, it is advisable not to rely on one clustering method to avoid overly optimistic or pessimistic inferences (Everitt et al. 2001). The hierarchical methods are often classified as *agglomerative* and *divisive*, of which the agglomerative approach has been the more popular and a large number of applications proves its practical utility (Everitt et al. 2001). Besides the quadratic memory and time requirements, the major drawback of the hierarchical clustering is the irreversibility of the fusions and divisions: Once two clusters have been fused, or one split, they cannot be later separated or merged (Everitt et al. 2001). The four methods studied here are agglomerative techniques which start from N clusters of individual objects and fuse pairs of clusters together until all the N objects are in the same cluster. Divisive methods operate in reverse direction by splitting the clusters until N single-member clusters have been produced. The hierarchical methods create a series of partitions $P_1, P_2, \ldots, P_N$, where P_i is a collection of i pairwise disjoint subsets $A \cap B = \emptyset$, where $A \subset P_i$, $B \subset P_i$, $P_i = \cup A_j$, and $1 \leq i \neq j \leq N$.

The agglomerative algorithms operate in a greedy and local manner. They merge at each stage a pair of clusters scored the best under some criterion which usually is a *proximity* (distance or similarity) measure. Therefore, the selection of an appropriate proximity measure is a central problem in applying the hierarchical methods. Measure unsuitable to the data may dramatically degrade the performance of a clustering method. The standard hierarchical methods store the proximities into a matrix and update them as clusters are fused. The iterative evaluation and updating of the pairwise cluster proximities requires at least $O(N^2)$ time, and the need to initially store proximities between all the objects explains the typical $O(N^2)$ space complexity. Proximity matrixes of the combinatorial methods may be updated without the original data using the Lance-Williams recurrence formula (Jain and Dubes 1988, Everitt et al. 2001). Only the proximity matrix is required to apply the formula in single and complete linkage, while to update the average linkage and Ward's matrixes the cluster sizes are needed besides the proximities.

The consecutive fusions are represented with a tree structure known as the *dendrogram* which is a mathematical and visual presentation of the clustering process (Jain and Dubes 1988, Everitt et al. 2001). The internal nodes represent clusters, the terminal nodes (leaves) the N objects, and the edges between the nodes indicate which clusters were fused to produce a new cluster. The dendrograms of the present study were weighed: Their height indicates the distances between the fused pairs. Figure 1 shows the progress of the single linkage method when vectors (1, 1), (1, 2), (3, 2), and (4, 1) were clustered with the Euclidean distance. For example, the last cluster is at height 2, because the smallest distance between clusters $A = \{(1, 1), (1, 2)\}$ and $B = \{(3, 2), (4, 1)\}$ was that between objects (1, 2) and (3, 2). The weighed dendrograms may also be inspected by considering only its levels, i.e. the order of the fusions. In Figure 1, levels 1, 2, and 3 refer to the first, second, and third fusion, respectively.

3.2. Choice of partition

Since searches utilizing the tree structure have not been found superior to simpler search strategies (El-Hamdouchi and Willett 1989), we decided to abandon the tree-based search and, instead, cut the trees at optimal height and used the resulting partitions to facilitate the cluster-based search. Selecting the optimal number of clusters is one of the central problems both in nonhierarchical and hierarchical cluster analysis (Everitt et al. 2001). A hierarchical cluster structure is converted into a partition by selecting one of the fusion or division levels of the structure, which is equivalent to "cutting" the dendrogram at a particular height. One

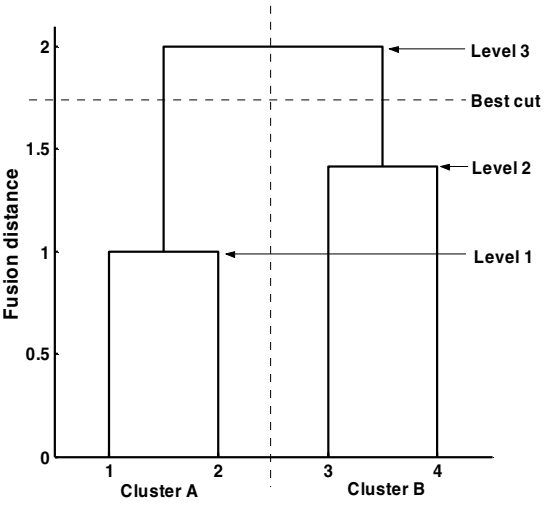

Fig. 1 Dendrogram produced by the single linkage method. The horizontal and vertical axes show the object labels and the fusion distances, respectively. According to the inconsistency coefficient, the best cut is between fusion levels 2 and 3. Thus, cutting the dendrogram at any height between $\sqrt{2}$ and 2 produces the optimal number of clusters, which is two

can visually inspect the dendrogram to find large changes in heights, which are taken to indicate the best level. This informal eyeballing approach, however, tends to be unfeasible with large data sets and is subject to *a priori* expectations. A number of formal methods to decide the best partition have been proposed, but their accuracy varies widely. Milligan and Cooper (1985) found in their comparison of 30 procedures that although the best procedures worked fairly well, the worst performers functioned only slightly better than chance level.

We applied the inconsistency (The Math Works Inc. 2002) and silhouette (Kaufman and Rousseeuw 1990) coefficients to the dendrograms created by the hierarchical methods. Since the partitions identified with the inconsistency coefficient were better than those found with the silhouette coefficient (see Section 5), only the former is described here. The inconsistency coefficient is similar to the well-known Mojena upper tail rule (Milligan and Cooper 1985, Everitt et al. 2001) which belonged to the best one-third in the comparison of Milligan and Cooper (1985). Both methods utilize statistics of the previous fusion levels, but the inconsistency coefficient considers only some of the previous levels, while the Mojena rule considers them all. The inconsistency coefficient for the ith ($i = 1, 2, \ldots, N-1$) fusion level α_i is

$$c_i = \frac{\alpha_i - \bar{\alpha}_z}{\sigma_z}, \tag{1}$$

where $\bar{\alpha}_z$ and σ_z are the respective mean and standard deviation of the heights of level α_i and the z highest fusion levels before it. It should be noted that the z heights are taken from the subtree rooted at the node of the ith level. Let the subtree contain l fusion levels. If $0 < l < z$, the l levels are considered, and when $l = 0$, the consistency coefficient is zero. Levels α_1 and α_{N-1} are the first and last merging levels, respectively. Cutting the dendrogram at any height between level $\max(c_i)$ and the level of the subtree immediately before $\max(c_i)$ yields the best partition with $N - i + 1$ clusters. Let us study Figure 1. Since $l = 0$ for subtrees of the two first levels, their coefficients are 0. If $z = 2$, the coefficient of the last level is $(2 - \bar{\alpha}_2)/\sigma_2 \approx 1.052$, where $\bar{\alpha}_2 \approx 1.4714$ and $\sigma_2 \approx 0.5024$ are the mean and standard deviation of heights 1, $\sqrt{2}$, and 2. Therefore, the optimal number of clusters is two and cutting the tree at height] $\sqrt{2}$, 2 [produces clusters A and B.

3.3. Principal components analysis

Principal components analysis (PCA) (Jolliffe 1986, Sharma 1996, Rencher 2002) is a multivariate statistical method that forms new variables (principal components) which are linear combinations of the original variables. The geometrical interpretation of PCA in an M-dimensional space is straightforward: the original M axes X are rotated to the direction of the maximum variance to create M new axes X'. The first new axis X'_1 is set in direction where the first principal component, i.e. projections of data points onto X'_1, accounts for the maximum of the total variance. Then, new axes are formed in the same manner using the variance which the previous principal components did not account for, until M new axes and their corresponding principal components have been created. The axes are orthogonal and the new variables uncorrelated with each other.

Popular approaches to facilitate PCA include finding the eigenstructure of the covariance matrix and solving the *singular value decomposition* (SVD) of the data set (Sharma 1996). Of these two approaches, SVD is computationally more efficient, because PCA is accomplished directly from the data without the need to build the covariance matrix. Both approaches require transformed data. Mean-corrected data is used, if it is assumed that the relative variances of the variables reflect their importance in forming new variables, while standardization of the original data allows weighing variables as equally important. The PCA solution is ideally such that the original variables have high correlations (loadings) with a small number of principal components, and few of the first principal components account for most of the variance of the data.

The PCA method may be seen as a tool that groups together correlated variables and, thus, helps the analysis of dependencies in data described with a large number variables. This analytical approach requires that one can successfully interpret the principal components using the loadings (Sharma 1996). On the other hand, PCA may be applied only to reduce the number of variables to facilitate analyses otherwise impossible due to processing time or memory limits. Dimension reduction is achieved by selecting $K (1 \leq K \ll M)$ largest principal components for further analysis. PCA is closely related to the *latent semantic indexing* (LSI) model which states that the ideas in a text are better captured by concepts than the index terms (Baeza-Yates and Ribeiro-Neto 1999, Belew 2000). Mapping of the document and query vectors into a lower dimensional space associated with the concepts is claimed to produce better retrieval performance than that of the original space. Dimension reduction is usually accomplished through the SVD technique. In this paper, we adopted purely the data reduction approach, and, thus, we neither attempt to interpret the principal components using the loadings, nor do we claim the improved retrieval levels as the LSI model does. Since it is known that dimension reduction often "polishes" data by removing some of its irregularities, such as noise, it is possible that the subspace may produce slightly better results than the original vector space.

3.4. Retrieval performance evaluation

The search capability of the cluster analysis methods was evaluated using the recall, precision, and effectiveness measures (Baeza-Yates and Ribeiro-Neto 1999, Belew 2000).

The effectiveness measure (E) combines recall (R) and precision (P) and allows their weighing with parameter $\beta > 0$:

$$E = 1 - \frac{(1 + \beta^2)PR}{\beta^2 P + R}. \tag{2}$$

Value of β indicates how many times more important recall is compared to precision. If $\beta = 1$, recall and precision are equally important. Recall is weighed more than precision, when $\beta > 1$ and vice versa. As the recall and precision measures, the effectiveness values are in [0, 1]. It should be noted that the effectiveness values are conversely interpreted to the recall and precision values. The small values indicate high effectiveness, while the large values are a sign of low effectiveness.

To calculate recall, precision, and the E-measure, the graded relevance assessment was binarized with two different approaches: (A) The relevant documents were weighed against the irrelevant ones, and (B) only the highly relevant documents were considered relevant, while all the other documents were labeled as irrelevant. In the following, we will name these relevance assessments in short as A and B, respectively. Statistical significance does not necessarily imply that the difference between methods is of practical importance (Keen 1992). Furthermore, relative differences only should not be reported, because they may be misleading. For these reasons, we use as a rule of thumb that the absolute differences in the results would be regarded as material if greater than 10%, noticeable if 5–10%, and not noticeable if smaller than 5% (Keen 1992).

4. Materials

4.1. The Finnish language and information retrieval

As one of the Finno-Ugrian languages, Finnish is very different from the Indo-European languages, such as the English language and the neighboring Indo-European languages, such as Swedish and Russian. Not surprisingly, the Finnish language has special characteristics which pose problems to the standard information retrieval methods developed for the processing of English. From the information retrieval viewpoint, the main problems are the various possibilities to inflect words and the abundance of compound words (Kekäläinen 1999, Alkula 2001). For example, each noun of the Finnish language has 15 cases, while English as only the nominative and genitive. Several different endings indicating number, case, possession, modality, tense, person, and other morphological characteristics, may be affixed to word stems. By considering all theoretically possible combinations, each substantive, adjective, and verb has 2,000, 6,000, and 12,000 inflected forms, respectively. These numbers are ten-fold larger when also all possible derivations are taken into account. Although in practice only some of the inflectional variants of each word are in a collection, it is obvious that indexing words without any preprocessing produces a large number of index entries and makes searching for any one of the different word forms demanding.

In highly agglutinative languages, such as Finnish or German, compound words are often spelled together. For example, the English compound word *nuclear power plant* is *ydinvoimalaitos* and *atomkraftwerk* in Finnish and German, respectively. The Dictionary of Modern Finnish contains approximately 130,000 compound words which represent two-thirds of all the words in the dictionary. Finnish compound words are difficult for information retrieval applications, because when included in the index as they are, searches using a part of a compound word are less likely to return documents containing the compound word. This problem is quite different from the challenge of phrase identification in English. It should be noted that because of the agglutinative nature of the Finnish language, the Finnish text documents may appear to be short. However, with an equal word count, a Finnish text tends

to be content-wise longer than an equal-length English one, because compounds are spelled together, and, furthermore, because Finnish lacks prepositions and articles.

4.2. The collection

The collection contained 53,893 articles published in three Finnish newspapers in 1988–1992 (Kekäläinen 1999). The average article length was 233 words and paragraphs were typically made of two or three sentences. The articles were mainly of domestic and foreign affairs and economics. The collection was processed with Finnish morphological analysis programs (Kekäläinen 1999, Alkula 2001). The inflected words were transformed into their morphological basic forms so that for each word its basic form was included into index. Some ambiguous words produced multiple basic form entities. The compound words were split into their parts, and the parts were further processed as the individual words. Splitting allows the use of any part of a compound word as search keys, and, consequently, improves recall. Some words, usually foreign proper names and typing errors, were not recognized because of the dictionary-based morphological analysis (Kekäläinen 1999, Alkula 2001).

The relevance of 16,539 documents to 30 queries had earlier been assessed with a four-level relevance scale (Kekäläinen 1999, Sormunen 2000, Kekäläinen and Järvelin 2002). These documents were judged as (0) irrelevant—the document was not about the topic of the query ($N = 14{,}588$), (1) marginally relevant—the topic was mentioned ($N = 886$), (2) fairly relevant—the topic was discussed briefly ($N = 700$), and (3) highly relevant—the topic was the main theme of the article ($N = 366$). The articles were assessed by two experienced journalists and two information retrieval specialists. The relevance of documents to 20 queries was evaluated by two or three persons and the rest 10 queries by one person. The assessors agreed in 73% of the assessments. Differences of one point (21%) were solved by taking the assessment from each judge in turn, and the most plausible grade after reevaluation was selected, when differences of two or three points (6%) were settled. The remaining 37,353 documents were marked as irrelevant.

The length of relevant documents at all relevance levels exceeded the average length of 233 words, but the highly relevant documents were, on average, shorter (306 words) than the fairly or marginally relevant documents whose respective average lengths were 314 and 334 words (Järvelin and Kekäläinen 2000). Highly relevant documents did not gain from higher document length in clustering because of the minor differences in average document lengths.

4.3. Numerical representation of text

The collection was represented as the $N \times M$ document-term matrix $\mathbf{D}$ for the clustering methods. As usual, the rows of the matrix correspond to the N documents (objects) which are characterized by the M terms (variables). The elements of each document vector $\mathbf{d}_i = (w_{i1}, w_{i2}, \ldots, w_{iM})$ are term weights computed according to the *tf·idf* scheme (e.g. Salton 1989):

$$w_{ij} = tf_{ij} \cdot (idf_j + 1) = tfi_j \cdot \left(\ln\left(\frac{N}{df_j}\right) + 1 \right), \tag{3}$$

where tf_{ij} is the frequency of the jth term in the ith document ($1 \le i \le N, 1 \le j \le M$), and 1 is added to keep the weight greater than zero, when a term occurs in all the documents. The matrix representation takes quadratic space, but was convenient for us for two reasons.

Firstly, the standard statistical and matrix computation software for clustering and dimension reduction assume observation matrix representation. Secondly, this representation facilitates the easy use of the conventional vector space model which is popular in document clustering (Willett 1988, Salton 1989).

A sample of 5,000 documents was studied, because preliminary runs showed that the available hardware and software could process roughly 10% of the data fast enough to allow series of quick experiments. The sample contained 1,926 relevant documents, because 26 relevant documents were accidentally missed while taking the sample. The irrelevant documents ($N = 3{,}074$) of the sample were randomly selected among the graded and non-graded ones, because their degree of irrelevance may differ. The documents graded as irrelevant were retrieved, and, consequently, resembled more the relevant documents than the non-graded irrelevant documents, which were not returned by the 30 queries. Therefore, using only graded irrelevant documents would have made the clustering task overly difficult, while clustering might have been too easy had solely non-graded irrelevant documents been used.

A pure random sampling would have been an inappropriate approach, because we especially aimed to study the graded relevance assessments in the clustering. The random sample might have excluded, for example, all documents highly relevant to some query, while keeping the marginally and fairly relevant documents. This is likely, because seven out of the 30 queries had less than six highly relevant documents and because there were clearly more marginally ($886/1{,}952 \cdot 100\% \approx 45\%$) and fairly (36%) relevant documents than highly relevant ones (19%). Moreover, some smaller queries might have been lost totally, because none of the relevant document would have been sampled. Therefore, to virtually make the study possible, we had to include all relevant documents in the sample.

Although a typical collection does not have as many relevant documents as the sample, the large proportion of the relevant documents does not necessarily mean that the retrieval would be markedly easier than in the full collection. For each query, we hoped that on average $1{,}952/30 \approx 65$ relevant documents would be brought together by the clustering methods. As in the full collection, this process is disturbed by two sets of documents: (A) roughly 1,900 documents that are relevant to the 29 other topics and (B) irrelevant documents, which are relevant to other topics than those in the current set of queries. The set A may be suspected to have some tendency to cluster away from the set of documents relevant to the topic under consideration, because they are relevant to another topic, which should reflect in their content, and because there were no topic overlaps. However, the document set A does not make the clustering task easier, quite the contrary. If one would start with set B and 65 relevant documents for topic S, and would then consider adding another set of 65 relevant documents for another topic T, the clustering task would not be easier than, but at least as difficult as before. The over 3,000 documents in the set B may be considered to be stronger random noise. If we assume the distribution of relevance levels similar in the sets A and B, the relevant documents in B are mainly marginally and fairly relevant because of the random sampling. These less relevant documents are prone to affect the clustering of the set A, because many the corresponding highly relevant documents were probably excluded from the sample.

The dimension reduction and clustering were performed with the Matlab software on an AlphaServer main frame of CSC—Scientific Computing Ltd. The computer had four 64-bit processors and 12 GB memory. Running times were from half an hour, taken by the clustering, to several hours needed to compute the silhouette coefficient. Since the main frame batched the runs, the actual running times were somewhat shorter. After the unrecognized words were excluded, the sample contained 53,621 index terms which had the total of 1,821,326 occurrences. The final $5{,}000 \times 13{,}693$ document-term matrix was produced with

further elimination of 773 words on a stoplist and 39,146 words that appeared in less than five documents. Preliminary tests with a smaller sample showed that exclusion of the most frequent and infrequent words had practically no effect on the clustering results. The matrix was constructed with Java programs written by the first author.

5. Results

5.1. Reduction and clustering of the sample

The mean-corrected and length-normalized sample was reduced with PCA which produced only 4,999 new non-zero variables. When the number of observations ($N = 5{,}000$) is smaller than the number of variables ($M = 13{,}693$), PCA creates $N - 1$ principal components, because a large number of columns of this type of matrix are linearly dependent (Horn and Johnson 1990). Unstandardized variables were used because of the same scale and equally-sized variances of the original variables. There are various methods to decide the number of principal components (Jolliffe 1986, Sharma 1996, Rencher 2002). One popular method is to plot the percentage of the variance accounted for each principal component (*eigenvalue*) against the component numbers and look for an "elbow". Figure 2 shows such *scree plot* (Jolliffe 1986, Sharma 1996, Rencher 2002) for our data. It can be seen that there is no apparent elbow and the largest principal components account only for a small amount of the total variance. This type of situation is quite common in real world data analyses. Here the sparse matrix made the results even further different from the ideal solution: The large number of zero entries explains the small equally-sized variances of both the original and new variables. Because of the missing elbow, we used another popular rule and selected the 1,500 largest principal components which accounted for 80% of the total variance (Jolliffe 1986, Rencher 2002). Through PCA we were able to

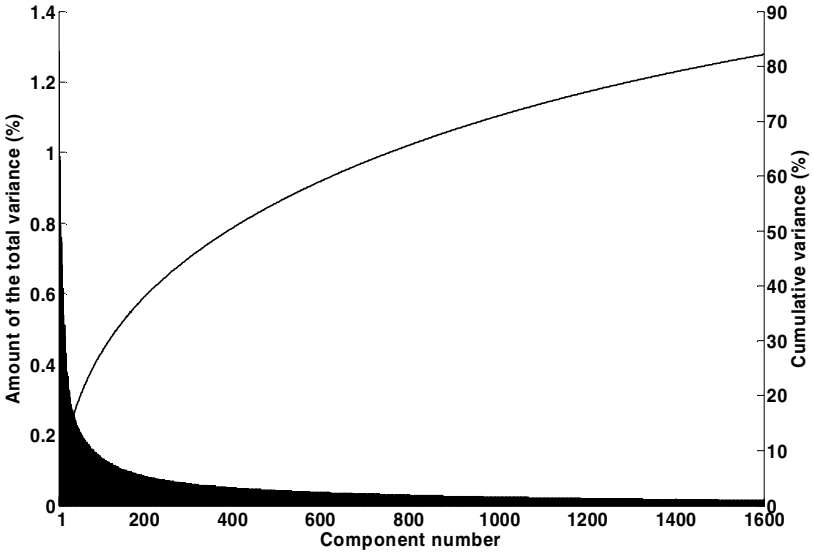

Fig. 2 Scree plot for the document sample. The first 1,600 out of the total 4,999 principal components plotted

conduct the hierarchical cluster analysis in space roughly nine times smaller than the original one.

To keep the clustering results of the original and reduced data as similar as possible, the clustering of new document vectors, i.e. the principal component score vectors, was performed using the Euclidean distance. Since mean-correction moves the data into origin, the angles between the vectors change, and, consequently, the cosine similarities between documents of the original and new data differ. If document vectors are normalized into the same length and the Euclidean distance is used in the clustering of the PCA results, the clustering results are exactly the same as those obtained with the original data and the cosine similarity measure, provided that the clustering method is invariant to order preserving proximity transformations. Korenius et al. (2004) discuss this matter in detail. Obviously, reduction of the data produces somewhat different result because part of information has been discarded.

The optimal partitions identified with the inconsistency coefficient for the single linkage, complete linkage, average linkage, and Ward's methods had 3197 ($z = 2$), 537 ($z = 2$), 668 ($z = 3$), and 160 ($z = 3$) clusters, respectively. When the respective dendrograms were cut at heights indicated by the silhouette coefficient, the partitions had 3652, 1755, 1455, and 1664 clusters. Although both methods had one maximum, several values nearly as large as the maxima made the solutions ambiguous. The coefficients agreed in single linkage, but for the other methods the inconsistency coefficient suggested levels closer to the root than the silhouette coefficient. Since reasonably large clusters are desirable, the solution given by the inconsistency coefficient was selected. Figure 3 presents the cluster size distributions, which were as expected. The single linkage method produced, due to the *chaining effect* (see e.g. Everitt et al. 2001), a few large and various singleton clusters, while the complete linkage method created many compact clusters. The average linkage method created more large clusters than the two former methods. The results of the Ward's method differ from those of the other clustering methods. This method discovered many equally-sized large clusters.

5.2. Best cluster performance

To study the clustering behavior, we evaluated clusters that contained one or more relevant documents. The typical recall values of the single and complete linkage clusters were extremely low (see Korenius et al. 2004). The subsequent tests showed that, as the typical complete linkage clusters, the typical average linkage and Ward's method clusters had also low precision. However, all the methods produced also large clusters with reasonably good recall and precision values. To study the best performance of clustering methods closer, we had to first decide how to measure the goodness of a cluster. This was a difficult task for two reasons. Firstly, the relative importance of recall and precision is known to be highly task dependent (Salton 1989). Secondly, there were a number of clusters too small to be considered as useful search results. Since the small clusters had typically high precision, we chose to use recall as the goodness measure, because recall is a more honest measure of performance than precision when small clusters abound. If precision had been used, tiny clusters would have been considered better than large ones, which are preferable in information retrieval.

Table 1 shows statistics for the best cluster of each query according to recall with relevance assessment B. Since the recall and precision measures were clearly better than the typical results, we decided to pursue the research further despite the poor typical results. It was also noteworthy that the best clusters contained fairly and marginally relevant documents in addition to the highly relevant ones. The complete linkage, average linkage, and

Table 1 Statistics describing the recall, precision, and effectiveness measures of the best cluster produced with the single linkage (SL), complete linkage (CL), average linkage (AL), and Ward's (WM) methods

Relevance level	Recall				Precision				Effectiveness ($\beta = 2$)			
Clustering method	SL	CL	AL	WM	SL	CL	AL	WM	SL	CL	AL	WM
Median	0.39	0.62	0.77	0.73	0.70	0.25	0.25	0.20	0.62	0.56	0.51	0.53
Mean	0.46	0.62	0.74	0.74	0.61	0.32	0.28	0.25	0.61	0.54	0.52	0.56
Standard deviation	0.27	0.23	0.25	0.25	0.38	0.20	0.21	0.21	0.21	0.17	0.19	0.18
Minimum	0.12	0.24	0.18	0.33	0.04	0.05	0.06	0.02	0.00	0.17	0.06	0.17
Maximum	1.00	1.00	1.00	1.00	1.00	0.89	0.94	0.94	0.89	0.82	0.81	0.89

The graded relevance assessment was binarized by considering only the highly relevant documents relevant and all the other documents irrelevant.

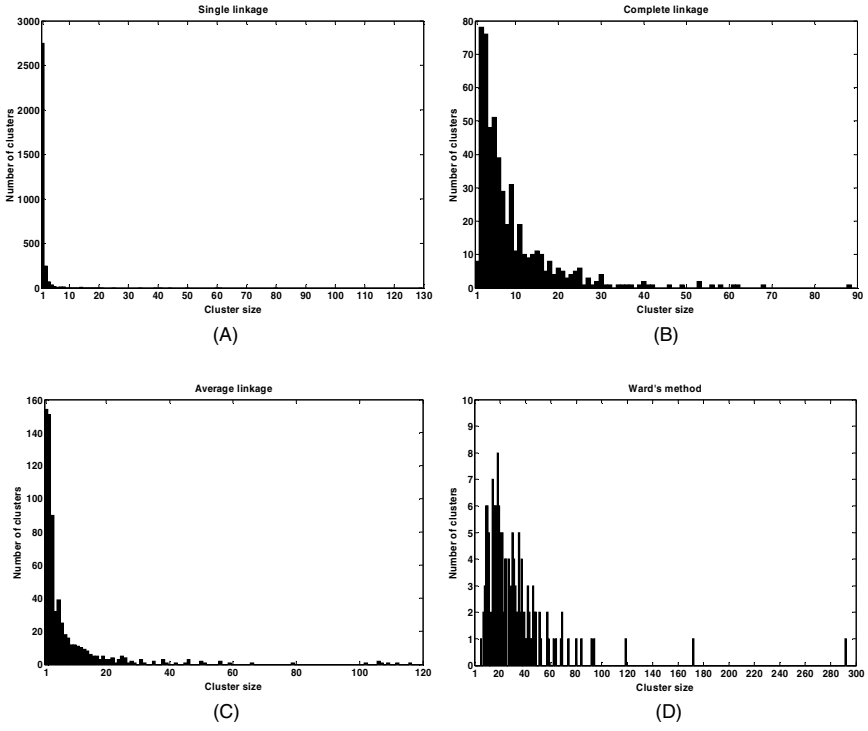

Fig. 3 Distribution of cluster sizes for the (A) single linkage, (B) complete linkage, (C) average linkage, and (D) Ward's clustering methods

Ward's methods produced high recall and low precision values, while the single linkage method had low recall and high precision. This discrepancy was detected also with the typical results (Korenius et al. 2004), because the single linkage characteristically produced tiny clusters.

5.3. Comparison of the cluster analysis and nearest neighbor search results

The results reported in the previous section were quite good, but unrealistic since the optimal cluster is not necessarily the cluster that matches the query best (Willett 1988). Therefore, we implemented a simple cluster-based search engine for the evaluation of the clustering solution. The cosine similarity between the query and the mean of each cluster was computed using the original data, and the cluster most similar to the query was returned to the user. This approach is even simpler than that of El-Hamdouchi and Willett (1989), because it does not make any use of the hierarchical cluster structure. We provide also nearest neighbor search results for the comparison purposes. Nearest neighbor (NN) is a well-known technique that allows the search and classification of objects (Mitchell 1997). The conventional exhaustive best-match search (Willett 1988), where documents are ranked in decreasing order by their similarity with the query, is in effect the popular NN search. The same number of documents as in the corresponding cluster was retrieved with the NN technique for each query.

Tables 2–5 depict statistics for the performance measures of the cluster-based and NN searches using relevance assessments A and B. The effectiveness measure is reported with $\beta = 0.5$ (precision twice as important as recall) and $\beta = 2$ (recall twice as important as precision). Since a large number of the documents retrieved with the cluster-based and NN searches were the same and the sample size ($N = 30$) was small, the statistical significance of differences was assessed with the two-tailed Wilcoxon signed ranks test (Pett 1997)—a nonparametric alternative of the paired t test. In the following, the results are studied mainly by comparing medians that are a more appropriate central statistic than means for skew distributions. Expectedly, the large differences in the median effectiveness showed the single linkage method clearly worse than the other methods. All the absolute differences, excluding those with relevance assessment B and $\beta = 0.5$, were large (21–47%) and statistically significant at $\alpha = 0.05$ ($p < 0.03$). Most of the differences between the complete linkage, average linkage, and Ward's methods were insignificant. However, with relevance assessment A and $\beta = 2$, the absolute differences of 21% and 23% were significant ($p < 0.002$) and favored the average and Ward's methods. Recall of both the clustering and NN methods was clearly lower with relevance assessment A than B, and precision of the both methods was clearly higher with assessment A than B.

Table 6 represents the absolute and relative differences in medians of the effectiveness measures of the NN and cluster-based searches. For example, the absolute and relative differences between the NN and Ward's method with relevance assessment A and $\beta = 0.5$ are $(0.42–0.34)\cdot 100\% \approx 8\%$ and $(0.42–0.34)/0.42\cdot 100\% \approx 19\%$, respectively. The NN searches were equal or better than the single and complete linkage searches and equal or worse than the average linkage and Ward's searches. Unfortunately, nearly all the differences were statistically insignificant. The NN and single linkage searches with relevance assessment A and $\beta = 2$ were significantly ($p < 0.05$) different even though the medians were the same. This result was an anomaly: Total of 21 value pairs was equal, and by chance eight of the extremely small ranked differences were for single linkage. As expected, the average linkage and Ward's methods performed better in comparison to the NN search than the single linkage method.

Table 7 shows the sizes the retrieved clusters and their ranks with assessments A and B. Ranks for each query were obtained by observing the position of the retrieved cluster in the order created by sorting the clusters in decreasing order according to the number of documents relevant to the query. The simple cluster-based search strategy identified the best cluster, i.e. cluster with the largest number of relevant documents, surprisingly well. The retrieved single linkage clusters were clearly smaller (median 4.00) than those of the other methods, and the retrieved complete linkage clusters (median 18.00) smaller than the retrieved average linkage

Table 2 Descriptive statistics of the recall, precision, and effectiveness (E) measures for the nearest neighbor (NN) and single linkage (SL) clustering based searches

	Recall		Precision		E ($\beta = 0.5$)		E ($\beta = 2$)	
Statistic	NN	SL	NN	SL	NN	SL	NN	SL
(A) The relevant documents versus the irrelevant documents								
Median	0.05	0.05	1.00	1.00	0.78	0.78	0.93	0.93
Mean	0.17	0.17	0.88	0.88	0.66	0.66	0.81	0.80
Standard deviation	0.21	0.22	0.30	0.31	0.31	0.32	0.23	0.24
Minimum	0.00	0.00	0.00	0.00	0.18	0.11	0.33	0.32
Maximum	0.64	0.64	1.00	1.00	1.00	1.00	1.00	1.00
(B) Highly relevant versus all the other documents								
Median	0.17	0.15	0.26	0.27	0.77	0.77	0.80	0.82
Mean	0.30	0.29	0.41	0.44	0.73	0.72	0.75	0.75
Standard deviation	0.33	0.34	0.38	0.43	0.23	0.27	0.25	0.27
Minimum	0.00	0.00	0.00	0.00	0.17	0.17	0.08	0.08
Maximum	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Graded relevance assessment was collapsed into binary one so that (A) all documents assessed as relevant were considered relevant and (B) only highly relevant documents were considered relevant and all the other documents irrelevant.

Table 3 Descriptive statistics of the recall, precision, and effectiveness (E) measures for the nearest neighbor (NN) and complete linkage (CL) clustering based searches

	Recall		Precision		E ($\beta = 0.5$)		E ($\beta = 2$)	
Statistic	NN	CL	NN	CL	NN	CL	NN	CL
(A) The relevant documents versus the irrelevant documents								
Median	0.27	0.27	0.88	0.90	0.40	0.41	0.69	0.69
Mean	0.34	0.34	0.81	0.81	0.43	0.43	0.63	0.63
Standard deviation	0.23	0.23	0.20	0.25	0.19	0.22	0.22	0.23
Minimum	0.02	0.02	0.08	0.08	0.15	0.05	0.13	0.10
Maximum	0.92	0.92	1.00	1.00	0.95	0.95	0.97	0.97
(B) Highly relevant versus all the other documents								
Median	0.50	0.50	0.27	0.21	0.69	0.78	0.60	0.61
Mean	0.55	0.51	0.28	0.26	0.71	0.73	0.59	0.62
Standard deviation	0.31	0.31	0.18	0.21	0.17	0.20	0.20	0.22
Minimum	0.00	0.00	0.00	0.00	0.24	0.18	0.21	0.20
Maximum	1.00	1.00	0.83	0.89	1.00	1.00	1.00	1.00

Graded relevance assessment was collapsed into binary ones so that (A) all documents assessed as relevant were considered relevant and (B) only highly relevant documents were considered relevant and all the other documents irrelevant.

Table 4 Descriptive statistics of the recall, precision, and effectiveness (E) measures for the nearest neighbor (NN) and average linkage (AL) clustering based searches

	Recall		Precision		E ($\beta = 0.5$)		E ($\beta = 2$)	
Statistic	NN	AL	NN	AL	NN	AL	NN	AL
(A) The relevant documents versus the irrelevant documents								
Median	0.45	0.48	0.81	0.87	0.38	0.33	0.51	0.48
Mean	0.47	0.49	0.75	0.78	0.39	0.37	0.51	0.49
Standard deviation	0.28	0.30	0.21	0.23	0.21	0.23	0.26	0.28
Minimum	0.02	0.00	0.06	0.00	0.11	0.06	0.18	0.11
Maximum	0.87	0.97	1.00	1.00	0.95	1.00	0.97	1.00
(B) Highly relevant versus all the other documents								
Median	0.67	0.78	0.17	0.18	0.79	0.79	0.55	0.50
Mean	0.67	0.69	0.24	0.25	0.74	0.73	0.58	0.55
Standard deviation	0.28	0.34	0.18	0.23	0.17	0.22	0.18	0.24
Minimum	0.00	0.00	0.00	0.00	0.22	0.17	0.27	0.06
Maximum	1.00	1.00	0.88	0.94	1.00	1.00	1.00	1.00

Graded relevance assessment was collapsed into binary ones so that (A) all documents assessed as relevant were considered relevant and (B) only highly relevant documents were considered relevant and all the other documents irrelevant.

Table 5 Descriptive statistics of the recall, precision, and effectiveness (E) measures for the nearest neighbor (NN) and Ward's method (WM) based searches

	Recall		Precision		E ($\beta = 0.5$)		E ($\beta = 2$)	
Statistic	NN	WM	NN	WM	NN	WM	NN	WM
(A) The relevant documents versus the irrelevant documents								
Median	0.46	0.52	0.82	0.83	0.42	0.34	0.51	0.46
Mean	0.48	0.51	0.72	0.75	0.39	0.36	0.51	0.48
Standard deviation	0.25	0.27	0.25	0.27	0.20	0.24	0.22	0.26
Maximum	1.00	0.89	1.00	1.00	0.96	1.00	0.98	1.00
Minimum	0.02	0.00	0.05	0.00	0.15	0.02	0.16	0.09
(B) Highly relevant versus all the other documents								
Median	0.62	0.82	0.17	0.18	0.80	0.79	0.62	0.55
Mean	0.61	0.71	0.21	0.22	0.78	0.77	0.62	0.60
Standard deviation	0.62	0.82	0.17	0.18	0.80	0.79	0.62	0.55
Maximum	1.00	1.00	0.88	0.94	1.00	1.00	1.00	1.00
Minimum	0.00	0.00	0.00	0.00	0.22	0.17	0.34	0.17

Graded relevance assessment was collapsed into binary ones so that (A) all documents assessed as relevant were considered relevant and (B) only highly relevant documents were considered relevant and all the other documents irrelevant.

Table 6 The absolute and relative differences (%) in medians of the effectiveness measures of the nearest neighbor and cluster-based searches

	$\beta = 0.5$				$\beta = 2$			
Difference	SL	CL	AL	WM	SL	CL	AL	WM
(A) The relevant documents versus the irrelevant documents.								
Absolute	0	−1	**5**	**8**	0	0	**3**	**5**
Relative	0	−3	**13**	**19**	0	0	**6**	**10**
(B) Highly relevant versus all the other documents.								
Absolute	0	−9	0	**1**	−2	−1	**5**	**7**
Relative	0	−13	0	**1**	−3	−2	**9**	**11**

Differences in favor of the single linkage (SL), complete linkage (CL), average linkage (AL), or Ward's (WM) methods are shown in bold font. Graded relevance assessment was collapsed into binary ones so that (A) all documents assessed as relevant were considered relevant and (B) only highly relevant documents were considered relevant and all the other documents irrelevant.

Table 7 Descriptive statistics (the five point summary, mean, and standard deviation) of the sizes of the retrieved clusters and their ranks, when the clusters were sorted in decreasing order according to the number relevant documents in the clusters

	Single linkage			Complete linkage			Average linkage			Ward's method		
Statistic	Size	Rank A	Rank B	Size	Rank A	Rank B	Size	Rank A	Rank B	Size	Rank A	Rank B
Minimum	1	1	1	4	1	1	3	1	1	14	1	1
Lower quartile	1.00	1.00	1.00	11.00	1.00	1.00	16.75	1.00	1.00	20.00	1.00	1.00
Median	4.00	1.00	1.00	18.00	1.00	1.00	24.50	1.00	1.00	31.50	1.00	1.00
Upper quartile	11.25	3.00	2.00	29.25	2.00	2.00	42.50	1.00	1.00	43.00	1.00	1.00
Maximum	61	5	3	88	5	3	112	3	2	92	5	3
Mean	8.47	2.07	1.35	22.80	1.60	1.43	37.83	1.24	1.19	36.17	1.31	1.14
Standard deviation	12.79	1.36	0.59	17.84	1.00	0.68	33.63	0.58	0.40	20.39	0.85	0.45
Valid N[a]	30	27	20	30	30	30	30	29	27	30	29	28

Graded relevance assessments were collapsed into binary ones so that (A) all documents assessed as relevant were considered relevant and (B) only highly relevant documents were considered relevant and all the other documents irrelevant. [a]Number of retrieved clusters that contained relevant documents.

and Ward's method clusters (medians 24.50 and 31.50). Many of the single linkage clusters did not contain any of the highly relevant documents. As with the best clusters, the relevant documents in the clusters were not only the highly relevant ones. The median percentage of the highly relevant documents in the retrieved single, complete, average linkage, and Ward's method clusters were 40, 25, 24, and 21%, respectively.

6. Discussion

A Finnish newspaper collection was clustered with four agglomerative hierarchical methods to facilitate cluster-based information retrieval. The collection was assessed into marginally, fairly, and highly relevant or irrelevant documents with respect to 30 queries. To allow application of the methods with the time complexity of $O(N^2)$, or worse, and the space complexity of $O(N^2)$, we focused on a sample of 5,000 documents that was further reduced utilizing principal components analysis (PCA). The reduced data were clustered with the single linkage, complete linkage, average linkage, and Ward's methods, whose dendrograms were cut at optimal heights identified with the inconsistency coefficient. The resulting partitions were analyzed by studying the retrieval performance with two binarized relevance assessments. The cluster-based searches were also compared to the nearest neighbor (NN) searches.

Analysis of the clusters containing relevant documents produced discouraging results (see Korenius et al. 2004). However, the best clusters were reasonably large and had good retrieval levels (see Table 1). The typical best complete linkage, average linkage, and Ward's cluster contained 62–77% of the relevant documents and 20–25% of their contents were relevant. The median effectiveness ($\beta = 2$) was 0.51–0.56. The imbalance between recall and precision was in expected direction, because recall was used as the goodness measure to take also into account the cluster size. The single linkage method behaved conversely to the other clustering methods studied here, because many of its clusters were expectedly small. Consequently, the best single linkage clusters had low recall, high precision, and median effectiveness of 0.62 which showed single linkage to be an outlying method. Although initial results reported in Korenius et al. (2004) were poor, the best clustering results, and the finding that the highly relevant documents tended to cluster together with the less relevant ones, provided evidence on behalf of the validity of the cluster hypothesis in the collection studied.

As pointed out by Willett (1988), the problem in concentrating on the optimal clusters is that there is no guarantee that they will be retrieved. Since the best cluster results were obviously too optimistic, we built a simple cluster-based search engine to evaluate the retrieval in a more realistic situation. When clusters were ranked according to their recall using assessments A and B, the retrieved best-matching clusters were always among the top five ranked clusters (see Table 7). Furthermore, most of the retrieved clusters were also the best clusters: More than 75% of the retrieved average linkage and Ward's method clusters were ranked the first. These clusters were quite large with the respective median sizes of 24.50 and 31.50 documents. It is possible that using recall as the measure of goodness introduced some bias to our study: The ranks of the retrieved clusters might have been lower, if precision had also been considered. However, the retrieved clusters were of surprisingly good quality bearing in mind the straightforward manner the cluster-based search was performed.

These results are in accord with the results of El-Hamdouchi and Willett (1989), who found that the partition utilizing search was more effective than the four tree-based search strategies. However, our approach to identify the optimal partition was not empirical but heuristic. To verify the decision based on the inconsistency coefficient, we evaluated systematically all partitions. Although better solutions were found, the selected partition was only slightly worse than the best partition. It seems that a simple heuristic, such as the inconsistency coefficient, is able to identify a solution that is feasible in information retrieval. Obviously, enumerative evaluation is not even possible with many real-world collections, which do not have relevance assessments.

As earlier, the single linkage results differed from those of the other clustering methods. The significant differences showed the single linkage searches 21–45% less effective than the searches based on the complete linkage, average linkage, and Ward's method results. Since

the retrieved clusters were also small (median size 4.00), the single linkage was undoubtedly the worst of the clustering methods. Single linkage it is not discussed further, because the earlier studies of Willett (1988) and El-Hamdouchi and Willett (1989) support this conclusion. Conversely to the results of Willett (1988), we found that the complete linkage method may be worse than the average linkage and Ward's methods. With relevance assessment A, the respective absolute differences in effectiveness were ($\beta = 0.5$) 8% and 7% and ($\beta = 2$) 21% and 23%, while with assessment B the differences were ($\beta = 0.5$) −1% and 1% and ($\beta = 2$) 11% and 6%. However, only the largest differences of 21% and 23% showed the two methods significantly better than complete linkage. The effectiveness of the average linkage and Ward's methods were better, because their recall was mostly clearly higher than that of complete linkage. The good performance of average linkage was consistent with the results of El-Hamdouchi and Willett (1989) which showed it the best of the four methods studied.

The comparison of the NN and cluster-based searches showed the two methods well-matched with a slight advantage to the searches facilitated by clustering. The NN searches were equal or better than searches based on the single and complete linkage clusters, and, on the other hand, the results were opposite when the performance of NN searches was compared to that of the average linkage and Ward's methods. Total six out of the 16 differences in medians were of noticeable size, five for and one against the two clustering methods, but none of them was statistically significant (see Table 6). The search results were evaluated with the binarized relevance assessments A and B (see Section 3.4). The former is the same as the traditional binary relevance assessment, where the degree of the relevance is not estimated. The latter assessment, in which only the highly relevant documents were judged relevant, allowed us to study how the highly relevant documents were retrieved.

Clustering had a bigger advantage over NN on relevance assessment A than B. When retrieval performance was evaluated with assessment A, the average linkage and Ward's methods were better than NN both when precision was weighed twice more than recall ($\beta = 0.5$) and recall was considered twice as important as precision ($\beta = 2$). The single and complete linkage methods were even with NN. These two clustering methods were worse than NN, when the results were inspected with assessment B. The average linkage and Ward's methods were still even with and better than NN with weighs $\beta = 0.5$ and $\beta = 2$, respectively. The results suggest that the partitions created with the average linkage and Ward's methods may offer better retrieval ability than the NN technique. Moreover, these methods may retrieve the highly relevant documents better than NN, if recall is more important than precision.

Although our results are not directly comparable with the partition search results of El-Hamdouchi and Willett (1989), the similarity of results suggest that when the characteristics of the Finnish language are taken into account, collections in Finnish, or in some other highly inflectional and agglutinative language, may be clustered as well as collections in English. The large number of inflected forms and compound words in Finnish were the main problems from information retrieval viewpoint. Transforming the inflected words into their morphological basic forms and splitting the compound words into their parts was an appropriate preprocessing approach to facilitate successful clustering of documents in Finnish.

Our results can also be used to partially validate the relevance assessments of the experts. Provided that the most relevant documents are truly highly relevant, it is likely that with assessment B many of the relevant documents are retrieved, but since the fairly and marginally relevant are considered irrelevant, also several irrelevant documents are retrieved. Consequently, recall and precision should be clearly different from those obtained when all relevant documents are clumped together. Since Tables 2–5 show consistently the recall of

assessment A lower than that of B and the precision of A higher than that of B, it seems that the assessors have done their work reliably.

The dimension reduction of data has been applied in information retrieval, but its use in document clustering has been infrequent. The results showed that PCA is a useful technique to enable the use of memory intensive hierarchical clustering. The number of terms could be reduced from 13,693 to 1,500, and, furthermore, the clustering of the greatly reduced data was successful when the pitfalls discussed by Korenius et al. (2004) were avoided. In our study, the lack of memory caused most problems. The running times of the dimension reduction and clustering methods were reasonable when a powerful computer was used, but the PCA method itself required large amounts of memory. However, there exist special algorithms, with smaller time and memory complexities than required by the algorithms used here, that utilize the sparsity of matrix to compute a partial SVD solution (Boley 1998). We plan to address these in the future research.

The standard hierarchical methods were an obvious choice, because the environments we are interested in are quite static and such where the accurate solutions are considered better than real-time responses. Since these methods have often been used in the document clustering, we were also able to compare our results to the results of the earlier studies. Moreover, the proven methods were, in our opinion, the best approach to tackle a novel problem. Future research should also consider the efficient new clustering algorithms presented in the area of data mining (Boley 1998, Nilsson 2002), because they might be used to generate clusters for search purposes. Besides supervised machine learning techniques, clustering might also be a tool to assess assessments: By comparing the relevance classes to a grouping, one can study to which degree the unsupervised method was able to recover the inherent structure (see, for example, Slonim and Tishby 2000).

Acknowledgments

This study was supported by the grants 50973 and 55243 from the Academy of Finland. The authors wish to thank the anonymous referees for many valuable comments which helped us to improve the paper.

References

Alkula R (2001) From plain character strings to meaningful words: Producing better full text databases for inflectional and compounding languages with morphological analysis software. Information Retrieval, 4(3-4):195–208.

Baeza-Yates R and Ribeiro-Neto B (1999) Modern Information Retrieval. ACM Press/Addison-Wesley, New York.

Belew RK (2000) Finding Out About. Cambridge University Press, Cambridge.

Boley D (1998) Principal direction divisive partitioning. Data Mining and Knowledge Discovery, 2(4):325–344.

Cutting DR, Karger DR, Pedersen JO and Tukey JW (1992) Scatter/Gather: A cluster-based approach to browsing large document collections. In Belkin N, Ingwersen P and Pejtersen AM, (eds.), Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, New York, pp. 318–329.

El-Hamdouchi A and Willett P (1989) Comparison of hierachic agglomerative clustering methods for document retrieval. The Computer Journal, 32(3):220–227.

Everitt BS, Landau S and Leese M (2001) Cluster Analysis, 4th edn. Arnold, London.

Horn RA and Johnson CR (1990) Matrix Analysis, 4th edn. Cambridge University Press, Cambridge.

Jain AK and Dubes RC (1988) Algorithms for Clustering Data. Prentice-Hall, New Jersey.

Jolliffe IT (1986) Principal Component Analysis. Springer-Verlag, New York.
Järvelin K and Kekäläinen J (2000) IR evaluation methods for retrieving highly relevant documents. In Belkin N, Ingwersen P and Leong MK, (eds.), Proceedings of the 23th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, New York, pp. 41–48.
Kaufman L and Rousseeuw PJ (1990) Finding Groups in Data. Wiley, New York.
Keen EM (1992) Presenting results of experimental retrieval comparisons. Information Processing & Management, 28(4):491–501.
Kekäläinen J (1999) The Effects of Query Complexity, Expansion and Structure on Retrieval Performance in Probabilistic Text Retrieval. Ph.D. Thesis, University of Tampere. Acta Universitatis Tamperensis, Vol. 678. URL: http://www.info.uta.fi/tutkimus/fire/archive/QCES.pdf
Kekäläinen J and Järvelin K (2002) Using graded relevance assessments in IR evaluation. Journal of the American Society for Information Science and Technology, 53(13):1120–1129.
Korenius T, Laurikkala J and Juhola M (2004) On applying the principal components analysis and cosine similarity for information retrieval. Manuscript available by a request from the authors.
Milligan GW and Cooper MC (1985) An examination of procedures for determining the number of clusters in a data set. Psychometrika, 50(2):159–179.
Mitchell TM (1997) Machine Learning. McGraw-Hill, New York.
Nilsson M (2002) Hierarchical clustering using non-greedy principal direction divisive partitioning. Information Retrieval, 5(4):311–321.
Pett MA (1997) Nonparametric Statistics for Health Care Research: Statistics for Small Samples and Unusual Distributions. Sage Publications, Thousand Oaks, California.
Rasmussen E (1992) Clustering algorithms. In Frakes W and Baeza-Yates R, eds. Information Retrieval: Data Structures and Algorithms. Prentice-Hall, Upper Saddle River, New Jersey, pp. 419–442.
Rencher AC (2002) Methods of Multivariate Analysis, 2nd edn. Wiley, New York.
Salton G (1983) Introduction to Modern Information Retrieval. McGraw-Hill, New York.
Salton G (1989) Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley, Reading, Massachusetts.
Sebastiani F (2002) Machine learning in automated text categorization. ACM Computing Surveys, 34(1):1–47.
Sharma S (1996) Applied Multivariate Techniques. Wiley, New York.
Slonim N and Tishby N (2000) Document clustering using word clusters via the information bottleneck method. In Yannakoudakis E, Belkin NJ, Leong M-K and Ingwersen P, eds. Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, New York, pp. 208–215.
Sormunen E (2000) A Method for Measuring Wide Range Performance of Boolean Queries in Full-Text Databases. Ph.D. Thesis, University of Tampere. Acta Universitatis Tamperensis, Vol. 748. URL: http://acta.uta.fi/pdf/951-44-4732-8.pdf
Sormunen E (2002) Liberal relevance criteria of TREC - counting on negligible documents? In Beaulieu M, Baeza-Yates R, Myaeng SH, Järvelin K, eds. Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, New York, pp. 324–330.
The Math Works Inc. (2002) Statistics Toolbox User's Guide, 4th edn. The Math Works Inc., Natick.
van Rijsbergen CJ (1980) Information Retrieval, 2nd edn. Butterworths, London.
Voorhees, E. (2001). Evaluation by Highly Relevant Documents. In Croft, WB, Harper, DJ, Kraft, DH & Zobel, J, eds. Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, New York. pp. 74–82.
Willett P (1988) Recent trends in hierarchic document clustering: A critical review. Information Processing & Management, 24(5):577–597.