The Curse of Dimensionality and Document Clustering

Giorgos Zervas and Stefan M Rüger
Department of Computing
Imperial College of Science, Technology and Medicine
180 Queen's Gate, London SW7 2BZ
s.rueger@doc.ic.ac.uk

Abstract

We suggest a document representation for clustering search engine results based on automatically generated keywords that are related to the query. We demonstrate the usefulness of this representation with a validation procedure using human relevance assessments of queries in a large document repository.

Introduction

Search Engines are a common gateway to huge document collections, be it the WWW or a collection of abstracts in an electronic book shop. It has been recognised that one limiting factor of search engine technology is the low precision of the results returned. It is not uncommon to get thousands or even millions of matches for a query such as "computer games". Even sophisticated ranking algorithms cannot know whether the user wants to browse documents on the latest advance in technology in this area or rather on entertaining products. We believe that cluster analysis of the documents that match a query provides a way to confront a user with different clusters/types of documents. Each cluster would either contain a high or a very low concentration of documents relevant to the user in accordance with the commonly accepted *cluster hypothesis* (van Rijsbergen 1979) for query results. This would allow users to quickly weed out whole clusters of irrelevant documents.

In the following we discuss three technical questions that arise in this context. How to represent documents to avoid the curse of dimensionality. How to efficiently (in roughly one second of CPU time) cluster 1000s of documents. How to validate the cluster hypothesis with the chosen representation and clustering algorithm. A visual-navigation search engine based on this work has been successfully implemented and is described elsewhere (Sewraz 1999).

1 The Curse of Dimensionality and Feature Reduction

A document collection can contain millions of different words. In our experiments with 528,155 US-American newspaper articles, we only kept nouns (based on Brill's tagger (Brill 1994)) with a medium document frequency: the noun had to appear in least 3 documents and in no more that 33% of all documents. Additionally, a small list of stop-words was used to eliminate obvious function words of the language. This resulted in a vocabulary of around 280,000 so-called *potentially interesting words*. A set H of documents returned by a query may still have a potentially-interesting-words vocabulary of 10,000s of different words. Consequently, the often-used word histogram representation of documents leads to high-dimensional vectors.

The problem with this kind of representations is that any two randomly picked vectors in a high-dimensional hypercube tend to have a constant distance from each other, no matter what the measure is! As an example, let $x, y \in [0,1]^n$ be drawn independently from a uniform distribution. The expectation value of their sum-norm distance is $n/3$ with a variance of $n/18$. For $n = 1,800$ (corresponding to a joint vocabulary of just 1,800 words for a word histogram representation) this means a typical distance of $|x - y|_1 = 600 \pm 10$. With increasing n the ratio between standard deviation and vector size gets ever smaller, as it scales with $1/\sqrt{n}$. Although word histogram document representations are by no means

random vectors, each additional dimension tends to not only spread the size of a cluster but also dilute the distance of two previously well-separated clusters. Hence, it seems prohibitive involving all semantic features (eg the words) of a document collection for document clustering.

Document clustering has attracted interest in the recent decades, eg (Salton 1968; Voorhees 1985; Rasmussen 1992), and much is known about the importance of feature reduction in general, eg (Krishnaiah and Kanal 1982), but little has been done to facilitate feature reduction for document clustering.

We suggest ranking the importance of each such word j with a weight

$$w_j = \frac{h_j}{d_j} \cdot h_j \log(|H|/h_j),$$

where h_j is the number of documents in H containing the word j, and d_j is the number of documents in the whole document collection D containing j. The second factor prefers medium matched-document frequency h_j, while the first factor prefers words that specifically occur in the matched documents. The highest-ranked words are meant to be related to the query. Indeed, we have "hardware", "software", "IBM" etc as the top-ranked words when querying for "computer". This seems to be a powerful approach to restrict the features of the matched documents to the top k ranked words, which we will call the *related words*. One important aspect is that the features are computed at query time. Hence, when above query is refined to "computer hardware", a completely new set of features would emerge automatically.

2 Clustering

We represent each matched document i as a k-dimensional vector v_i, where the j-th component v_{ij} is a function of the number of occurrences t_{ij} of the j-th ranked related word in the document i:

$$v_{ij} = \log_2(1 + t_{ij})$$

We project the vector v_i onto the k-dimensional unit sphere obtaining a normalised vector u_i that represents the document i. We deem the Euclidean distance between u_a and u_b a sensible *semantic distance* between two documents a and b in the document subset H returned by a query with respect to the complete document collection D.

We use a standard iterative clustering technique to compute N clusters of documents. The N seeds for the initial cluster centres are obtained by a full hierarchical clustering of the best-ranked 100 documents resulting from the query.

3 Validation

Any clustering method — even random assignment — leads to a partition of the documents. We propose a method to assess the quality of the clustering process based on human-expert data. We have used the 1997-1998 collection of the TREC data (Voorhees and Harman 1999) with 528,155 documents, mainly newspaper articles, 100 queries and corresponding relevance assessments. We ignored queries that contained fewer than 40 relevant documents and divided the remaining 61 queries randomly into test data (15) and training data (46). For each of the training queries we would run the query in a standard search engine and partition the set H of 1000 best-ranked documents into 6 clusters $H_1, \ldots, H_6$ according to above scheme. A certain proportion p_* of the documents in H would be relevant according to the human relevance assessments. A clustering, where the proportions $p_1, \ldots, p_6$ of relevant documents in each cluster, respectively, were either 0 or 1 would be ideal. Contrary, a clustering where $p_1 = \ldots = p_6 = p_*$, could not assist a human user at all to quickly weed out big sets of irrelevant documents (see the figure below). We came up with a quality measure that assigns 1 in the best case and 0 in the worst case and linearly interpolates between these cases. Averaging over all available training queries gives insight into the typical behaviour of a clustering routine in the context of a particular search engine.

4 Experiment Results and Conclusions

The experimental procedure was applied to queries with different dimensions (8, 16, 22, 28, 32, 36, 50, 64, 96, 128, 256, 512, 1024) for the document representation and three linkage methods each (single, average and complete) for the seeding of the iterative-clustering centres. For each of the combinations, we computed a cluster quality as described above and averaged this quality over the training queries.

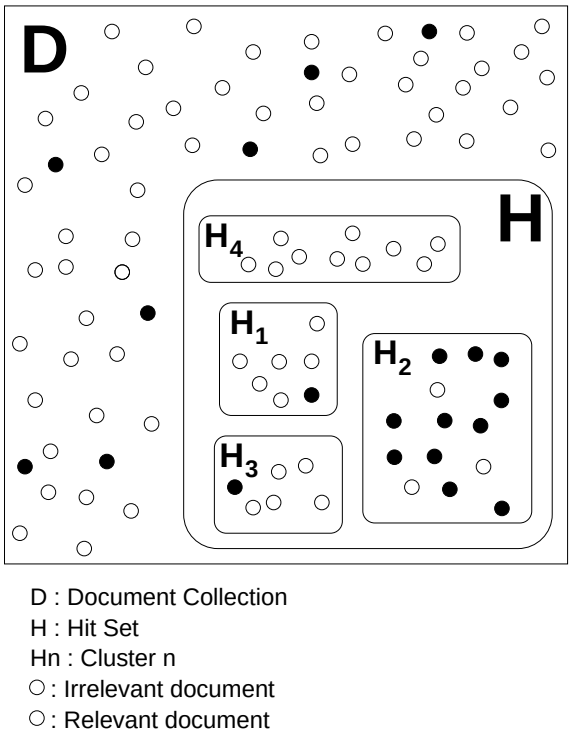

Figure 1: The document collection after searching and clustering

Our studies confirmed that the average cluster quality is significantly higher than random cluster assignments, which supports the cluster hypothesis. We were also able to tune the parameters of the clustering procedure, most notably the number k of features used in the document representation. Our preliminary findings indicate that $k \approx 30$ is sufficient for good clustering results, and that single linkage for the hierarchical clustering seeds seems to outperform the other linkage methods. All our findings have been confirmed on the test queries.

Bibliography

References

Brill, E. (1994). Some advances in rule-based part of speech tagging. In *AAAI*.

Krishnaiah, P. R. and L. N. Kanal (Eds) (1982). *Handbook of Statistics: Classification, Pattern Recognition and Reduction of Dimensionality*, Volume 2. North-Holland Publishing Company.

Rasmussen, E. (1992). Clustering algorithms. In W. B. Frakes and R. Baeza-Yates (Eds), *Information Retrieval: Data Structures and Algorithms*, pp. 419–442. Prentice Hall.

van Rijsbergen, C. J. (1979). *Information Retrieval* (2nd ed). London: Butterworth.

Salton, G. (1968). *Automatic information organization and retrieval*. New York: McGraw-Hill.

Sewraz, S. (1999). *A Visual Information-Retrieval Navigator*. MSc Thesis, Imperial College. For an extended abstract, see http://www.doc.ic.ac.uk/~smr3/pub/virn.html.

Voorhees, E. (1985). The cluster hypothesis revisited. In *Proceedings of ACM SIGIR*, pp. 188–196.

Voorhees, E. M. and D. K. Harman (1999). *Information Technology: The Seventh Text REtrieval Conference (TREC-7)*. NIST. http://trec.nist.gov.

Acknowledgements: This work was partly supported by the Fujitsu European Centre for Information Technology (FECIT).