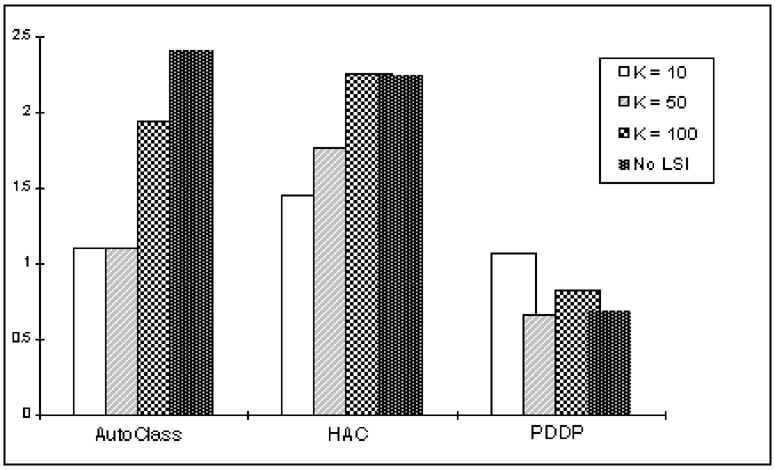

Figure 7: Comparison of Entropies for the E_1, With and Without LSI.

suggest), from algorithm to algorithm.

Also, both HAC and AutoClass ran dramatically faster on the data with small number of dimensions than on the original data. For the small data set (E1) used for these experiments, the time to run SVD was relatively small. Hence, the overall runtime including SVD computation and clustering improved. However, for a large matrix, the runtime for computing SVD could be quite substantial and may make it impractical to perform dimensionality reduction before clustering.

The precise cost of the SVD computation for both LSI and PDDP depends on many variables, including the number of eigenvalues sought and their distribution, and hence is much beyond the scope of this paper. The underlying Lanczos iteration needed to obtain the $k \gg 1$ eigenvalues in LSI is more or less the same as that used to obtain the single eigenvalue in PDDP, except that in the latter there is no need to deflate eigenvalues as they converge to prevent them from re-appearing. However, at least k, and often at least $2k$, iterations are needed in the LSI case [Ber92]. On the other hand, obtaining the single leading eigenvector can be accomplished in relatively few iterations, typically around 15 to 20, and never more than 25 in our application. We also note that though the SVD computation in PDDP must be repeated, it is applied to the entire matrix only once.

Finally, we note that we have not run ARHP on the LSI datasets, as ARHP would have required the computation of frequent item sets (documents) based on a very small number of (k) transactions. Frequent item sets discoverd using a small number of transactions are not likely to be useful, because they are not statistically significant. Hence, the hypergraph constructed from these frequent item sets does not capture the original relationships present among the documents.

5.2 Scalability of Clustering Algorithms

The scalability of our clustering methods is essential if they are to be practical for large numbers of documents. In this section we give some experimental evidence that our methods are indeed scalable.

We have applied our clustering methods to a large data set denoted "D1" consisting of 2,340 documents using a dictionary of 21,839 words. We then constructed three other data sets labeled D3, D9, D10 using reduced dictionaries using the same strategies as E3, E9, E10, respectively, in Table 1. The number of words in the resulting reduced dictionaries is reported in Fig. 9.

Scalability of PDDP

Figures 8 and 9 illustrate the performance of the PDDP algorithm on these datasets. Fig. 8 shows that the entropies for D1 and D10 are lower than those for D3 and D9, when the number of clusters agrees. This is consistent with the results shown in Fig. 6.

The PDDP algorithm is based on an efficient method for computing the principal direction. The method for computing the principal direction is itself based on the Lanczos method [GV96] in which the major cost arises from matrix-vector products involving the term frequency matrix. The total cost is the cost of each matrix-vector product times the total number of products. The number of products has never exceeded 20 in any of the experiments we have tried. The term frequency matrix is a very sparse matrix: in the D1 data set only about 0.68% of the entries are nonzero. The cost of a matrix-vector product involving a very sparse matrix depends not on the dimensions of the matrix, but rather

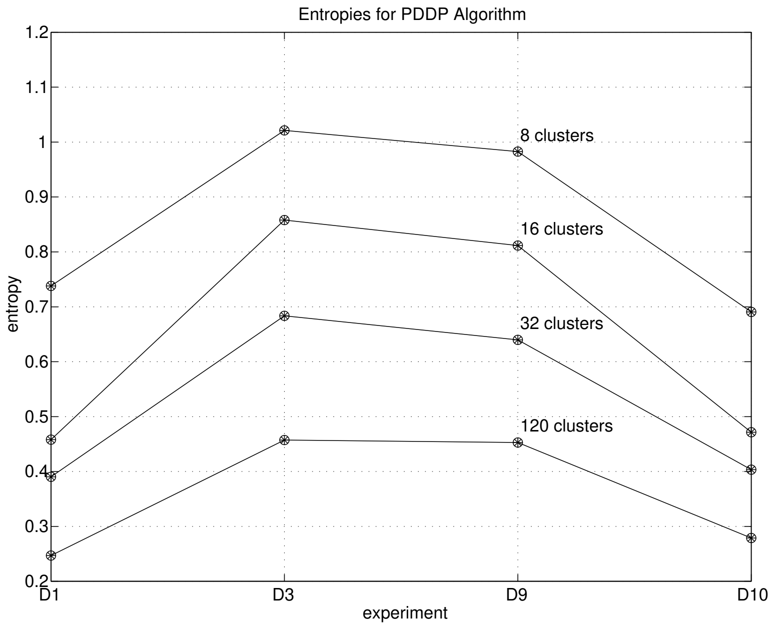

Figure 8: Entropies from the PDDP algorithm with various number of clusters.

on the number of nonzero entries. Using the notation of the previous section, we seek the leading singular vector of the dense matrix $(\widehat{M} - \mathbf{ce})$, but matrix-vector products involving this matrix can be computed as shown in equation (1) without forming this matrix explicitly. This is all that is required for the Lanczos method. Hence the total cost of the PDDP algorithm is governed by the number of nonzero entries in the term frequency matrix. This is illustrated in Fig. 9. Notice that even though the data set D10 has many fewer words than the other data sets, its cost is more than for D3 or D9 because D10's term frequency matrix is 10 times more dense than D1's matrix: about 6.9% of its entries are nonzero.

Scalability of ARHP

The problem of finding association rules that meet a minimum support criterion has been shown to be linearly scalable with respect to the number of transactions [AMS$^+$96]. It has also been shown in [AMS$^+$96] that association rule algorithms are scalable with respect to the number of items assuming the average size of transactions is fixed. Highly efficient algorithms such as Apriori are able to quickly find association rules in very large databases

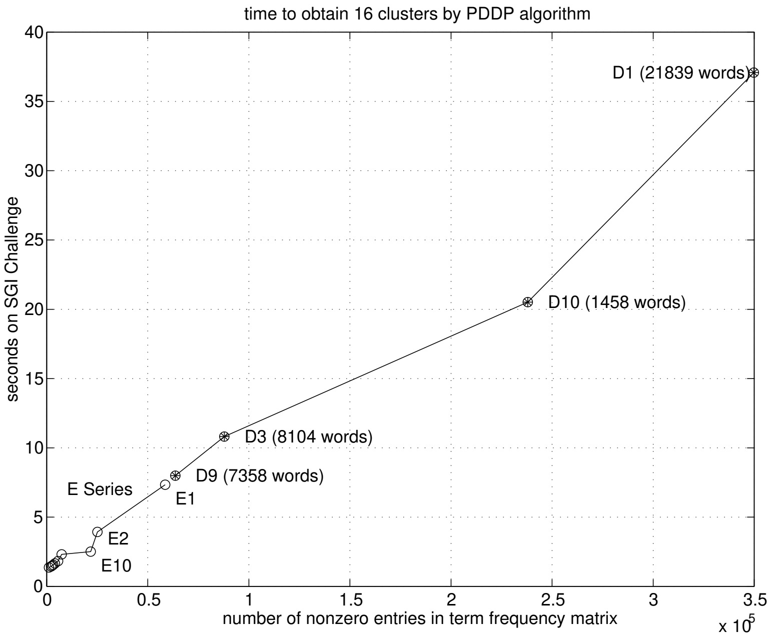

Figure 9: Times for the PDDP algorithm on an SGI versus number of nonzeros in term frequency matrix, for both E and D series with 16 clusters. The D experiments had 2,340 documents and the indicated number of words, and were constructed just like the corresponding E experiments in Table 1.

provided the support is high enough.

The complexity of HMETIS for a k-way partitioning is $O((V + E) \log k)$ where V is the number of vertices and E is the number of edges. The number of vertices in an association-rule hypergraph is the same as the number of documents to be clustered. The number of hyperedges is the same as the number of frequent item-sets with support greater than the specified minimum support. Note that the number of frequent item sets (i.e., hyperedges) does not increase as the number of words increases. Hence, our clustering method is linearly scalable with respect to the number of words in the documents.

Figure 10 shows the entropies from the ARHP algorithm. This result is also consistent with Figure 6 where the entropies for D1 and D10 are lower than for D3 and D9. Each data set produces different size hypergraph. Table 2 shows the size of hypergraph for several experiments from E and D series. Figure 11 shows the CPU time for partitioning these

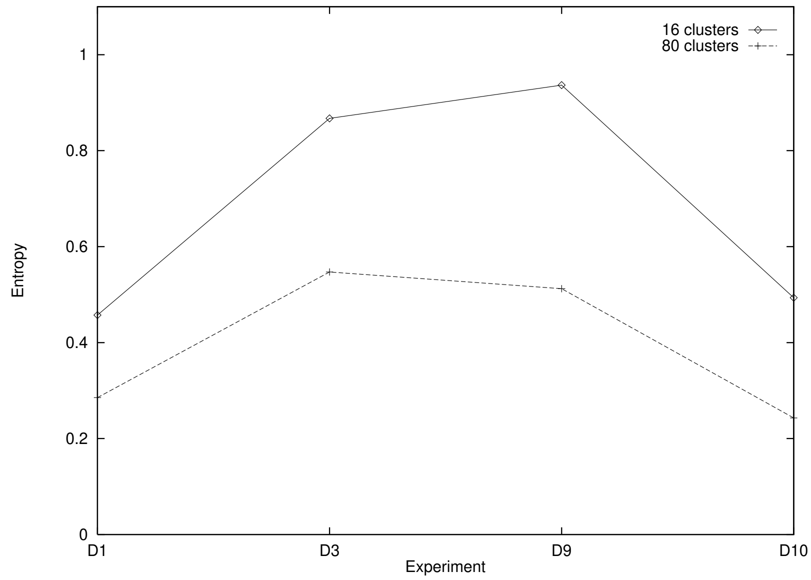

Figure 10: Entropies from the ARHP algorithm with various number of clusters.

hypergraphs. The run time for partitioning these hypergraphs supports the complexity analysis that says the run time is proportional to the size of the hypergraph $(V + E)$.

6 Search for and Categorization of Similar Documents

One of the main tasks of the agent is to search the Web for documents that are related to the clusters of documents. The key question here is how to find a representative set of words that can be used in a Web search. With a single document, the words appearing in the document become a representative set. However, this set of words cannot be used directly in a search because it excessively restricts the set of documents to be searched. The logical choice for relaxing the search criteria is to select words that are very frequent in the document.

The characteristic words of a cluster of documents are the ones that have high document frequency and high average text frequency. Document frequency of a word refers to the frequency of the word across documents. Text frequency of a word refers to word frequency within a document. We define the TF word list as the list of k words that have the highest average text frequency and the DF word list as the list of k words that have the highest

Experiments	Number of Edges	Number of Vertices
E1	12091	185
E3	6572	185
E9	4875	185
E10	15203	185
D1	95882	2068
D3	81677	2065
D9	128822	2028
D10	89173	2147

Table 2: Size of hypergraphs for several experiments from E and D series.

document frequency.

For each cluster, the word lists TF and DF are constructed. $TF \cap DF$ represents the characteristic set of words for the cluster, as it has the words that are frequent across the document and have high average frequency. The query can be formed as

$$(c_1 \wedge c_2 \ldots \wedge c_m) \wedge (t_1 \vee t_2 \ldots \vee t_n)$$

where $c_i \in TF \cap DF$ and $t_i \in TF - DF$.

We formed queries from the business capital cluster discussed in Section 5. We found the characteristic words of the cluster ($TF \cap DF$) and issued the following query to Yahoo web search engine:

```
+capit* +busi* +financ* +provid* +fund* +develop* +compani* +financi* +manag*
```

The search returned 2280 business related documents. We then added the most frequent words that were not in the previous list ($TF - DF$) to form the following query:

```
+capit* +busi* +financ* +provid* +fund* +develop* +compani* +financi* +manag*
loan* invest* program* credit* industri* tax* increas* cost* technologi*
sba* project*
```

AltaVista search using this query returned only 372 business related documents which seemed highly related to the existing documents in the cluster. First page returned by the query is shown in Figure 12.

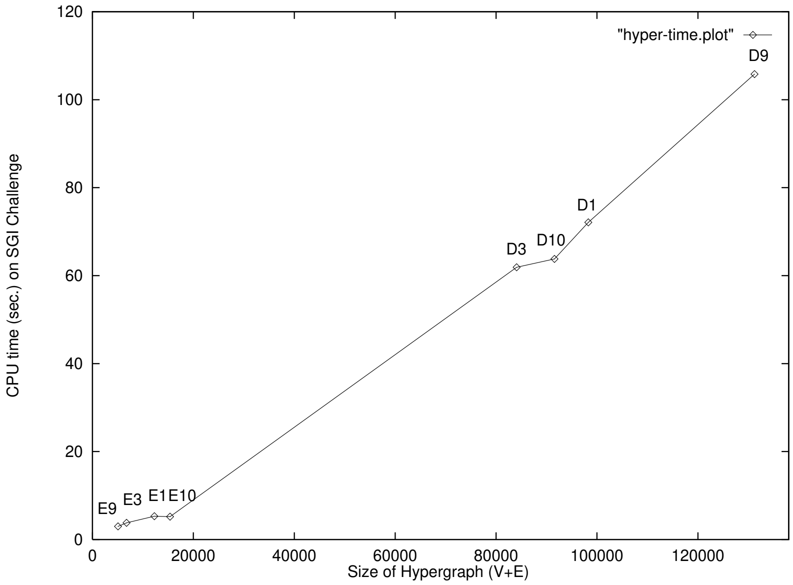

Figure 11: Hypergraph partitioning time for both E and D series. The number of partitions was about 16.

The documents returned as the result of queries can be handled in several ways as shown in Figure 1. ARHP could be used to filter out non-relevant documents among the set of documented returned by the query as discussed in Section 4.1. The degree of filtering can be increased either by setting higher support criteria for association rules discovery or by having a tighter connectivity constraint in the partition.

Resulting documents can be incrementally added to the existing clusters using ARHP or PDDP depending on the method used for clustering. With ARHP, for each new document, existing hyperedges are extended to include the new document and their weights are calculated. For each cluster, the connectivity of this new document to the cluster is measured by adding the weights of all the extended hyperedges within the cluster. The new document is placed into the cluster with the highest connectivity. The connectivity ratio between the chosen cluster and the remaining clusters indicates whether the new document strongly belongs to the chosen cluster. If the connectivity of the document is below some threshold for all clusters, then the document can be considered as not belonging to any of the cluster.

With PDDP, the binary tree can also be used to filter new incoming documents by placing

YAHOO! SEARCH — Powered by ALTAVISTA

MLB Playoffs - NHL Preseason -- Drop-Off Locator

Categories - Sites - **Alta Vista Web Pages** | Net Events - Headlines - Amazon.com Related Books

Alta Vista Web Pages (1-20 of 372)

- SBA Loans Take A New Direction - SBA Loans Take A New Direction. April, 1993. While the restrictive conditions in the commercial lending environment show some signs of abating, obtaining..
  --http://www.ffgroup.com/contractor/sba_8a/504loans.html

- SBA: Small Business Act of 1958 and PL 104-208, Approved 9/30/96 - This compilation includes PL 104-208, approved 9/30/96. SMALL BUSINESS INVESTMENT ACT OF 1958. (Public Law 85-699, as amended) Sec. 101. SHORT TITLE This..
  --http://www.sbaonline.sba.gov/INV/sbaact.html

- New Haven Enterprise Community Summary - EC Summary Contact EC Summary Maps. STRATEGIC PLAN SUMMARY. Introduction. The New Haven Enterprise Community Strategic Plan marshals our community's...
  --http://www.hud.gov/cpd/ezec/ct/ctnewhav.html

- ABIOGENESIS SOFTWARE - Business Venture Finance Investment Info - The Abiogenesis Business Finance Resource Site. Abiogenesis provides software for the creation of computer dictionaries. Setting up and capitalizing your..
  --http://www.abiogenesis.com/AbioDocs/Finance.html

- Financial Information - Finance Executives. General Web Resources. CorpFiNet. An Introduction to the WWW for Executives. SuperCFOs. Well-written article from Fortune discusses...
  --http://www.unf.edu/students2/jroger2/finance.html

- Fairfax County Business Services and Resources (Part 3) - Business Services and Resources. Arts. Associations. Career Development/Continuing Education. Child Care. Chambers of Commerce and Other Business...
  --http://www.eda.co.fairfax.va.us/fceda/do_bus/b_resrc3.html__28506-4

- Canadian Financial Regulation: A System in Transition - Commentary 78; Financial Regulation March 19, 1996. Canadian Financial Regulation: A System in Transition. by Edwin H. Neave. Abstract. Planned revisions..
  --http://www.cdhowe.org/eng/word/word-5.html

- FBS | Business Page | re:BUSINESS | Summer 1996 - RE: Business. SUMMER 1996. THE FIRST AMERICAN 401(K) SOLUTION FOR EMPLOYEES' RETIREMENT. Today, many businesses are setting up 401(k) plans for...
  --http://www.fbs.com/biz_pages/newsletters/96summer.html

- CPB TV Future Fund Business Plans - TV Future Fund. Business Plan Outline. Updated

Figure 12: Search results from Yahoo.

the document on one or the other side of the root hyperplane, then placing it on one or the other side the next appropriate hyperplane, letting it percolate down the tree until it reaches a leaf node. This identifies the cluster in the original tree most closely related to the new incoming document. If the combined scatter value for that cluster with the new document is above a given threshold, then the new document is only loosely related to that cluster, which can then be split in two.

7 Conclusion

In this paper we have proposed an agent to explore the Web, categorizing the results and then using those automatically generated categories to further explore the Web. We have presented sample performance results for the categorization (clustering) component, and given some examples showing how those categories are used to return to the Web for further exploration.

For the categorization component, our experiments have shown that the ARHP algorithm and the PDDP algorithm are capable of extracting higher quality clusters while operating much faster compared to more classical algorithms such as HAC or AutoClass. This is consistent with our previous results [MHB+97]. The ARHP algorithm is also capable of filtering out documents by setting a support threshold. Our experiments show that the PDDP and ARHP algorithms are fast and scale with the number of words in the documents.

To search for similar documents keyword queries are formed by extending the characteristic word sets for each cluster. Our experiments show that this method is capable of producing small sets of relevant documents using standard search engines.

In the future, we will explore the performance of the entire agent as an integrated and fully automated system, comparing the relative merits of the various algorithms for clustering, query generation, and document filtering, when used as the key components for this agent. In particular, we will conduct further experimental evaluation of query generation mechanism and classification of new documents into existing clusters. Another area for future work involves the development of a method for evaluating quality of clusters which is not based on *a priori* class labels.

References

[Ack97] M. Ackerman et al. Learning probabilistic user profiles. *AI Magazine*, 18(2):47–56, 1997.

[AFJM95] R. Armstrong, D. Freitag, T. Joachims, and T. Mitchell. WebWatcher: A learning apprentice for the world wide web. In *Proc. AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments*. AAAI Press, 1995.

[AMS+96] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A.I. Verkamo. Fast discovery of association rules. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smith, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 307–328. AAAI/MIT Press, 1996.

[And54] T.W. Anderson. On estimation of parameters in latent structure analysis. *Psychometrika*, 19:1–10, 1954.

[BDO95] M. W. Berry, S. T. Dumais, and Gavin W. O'Brien. Using linear algebra for intelligent information retrieval. *SIAM Review*, 37:573–595, 1995.

[Ber76] C. Berge. *Graphs and Hypergraphs*. American Elsevier, 1976.

[Ber92] M. W. Berry. Large-scale sparse singular value computations. *International Journal of Supercomputer Applications*, 6(1):13–49, 1992.

[Bol97] D.L. Boley. Principal Direction Divisive Partitioning. Technical Report TR-97-056, Department of Computer Science, University of Minnesota, Minneapolis, 1997.

[BSY95] Marko Balabanovic, Yoav Shoham, and Yeogirl Yun. An adaptive agent for automated Web browsing. *Journal of Visual Communication and Image Representation*, 6(4), 1995.

[CS96] P. Cheeseman and J. Stutz. Baysian classification (autoclass): Theory and results. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smith, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 153–180. AAAI/MIT Press, 1996.

[DDF+90] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R Harshman. Indexing by latent semantic analysis. *J. Amer. Soc. Inform. Sci.*, 41:41, 1990.

[DEW96] R. B. Doorenbos, O. Etzioni, and D. S. Weld. A scalable comparison shopping agent for the World Wide Web. Technical Report 96-01-03, University of Washington, Dept. of Computer Science and Engineering, 1996.

[DH73] Richard O. Duda and Peter E. Hart. *Pattern Classification and scene analysis.* John Wiley & Sons, 1973.

[FBY92] W. B. Frakes and R. Baeza-Yates. *Information Retrieval Data Structures and Algorithms.* Prentice Hall, Englewood Cliffs, NJ, 1992.

[Fra92] W. B. Frakes. Stemming algorithms. In W. B. Frakes and R. Baeza-Yates, editors, *Information Retrieval Data Structures and Algorithms*, pages 131–160. Prentice Hall, 1992.

[GV96] G. H. Golub and C. F. Van Loan. *Matrix Computations.* Johns Hopkins Univ. Press, 3rd edition, 1996.

[HBML95] K. Hammond, R. Burke, C. Martin, and S. Lytinen. FAQ-Finder: A case-based approach to knowledge navigation. In *Working Notes of the AAAI Spring Symposium: Information Gathering from Heterogeneous, Distributed Environments.* AAAI Press, 1995.

[HKKM97a] E.H. Han, G. Karypis, V. Kumar, and B. Mobasher. Clustering based on association rule hypergraphs (position paper). In *Workshop on Research Issues on Data Mining and Knowledge Discovery*, pages 9–13, Tucson, Arizona, 1997.

[HKKM97b] E.H. Han, G. Karypis, V. Kumar, and B. Mobasher. Clustering in a high-dimensional space using hypergraph models. Technical Report TR-97-063, Department of Computer Science, University of Minnesota, Minneapolis, 1997.

[HKKM98] E.H. Han, G. Karypis, V. Kumar, and B. Mobasher. Hypergraph based clustering in high-dimensional data sets: A summary of results. *Bulletin of the Technical Committee on Data Engineering*, 21(1), 1998.

[Jac91] J. E. Jackson. *A User's Guide To Principal Components.* John Wiley & Sons, 1991.

[JD88] A.K. Jain and R. C. Dubes. *Algorithms for Clustering Data.* Prentice Hall, 1988.

[KAKS97] G. Karypis, R. Aggarwal, V. Kumar, and S. Shekhar. Multilevel hypergraph partitioning: Application in VLSI domain. In *Proceedings ACM/IEEE Design Automation Conference*, 1997.

[KLSS95] T. Kirk, A. Y. Levy, Y. Sagiv, and D. Srivastava. The information manifold. In *Working Notes of the AAAI Spring Symposium: Information Gathering from Heterogeneous, Distributed Environments.* AAAI Press, 1995.

[Koh88] T. Kohonen. *Self-Organization and Associated Memory.* Springer-Verlag, 1988.

[KW96] C. Kwok and D. Weld. Planning to gather information. In *Proc. 14th National Conference on AI*, 1996.

[LF78] S.Y. Lu and K.S. Fu. A sentence-to-sentence clustering procedure for pattern analysis. *IEEE Transactions on Systems, Man and Cybernetics*, 8:381–389, 1978.

[LS97] H. Vernon Leighton and J. Srivastava. *Precision among WWW search services (search engines): Alta Vista, Excite, Hotbot, Infoseek, Lycos.* http://www.winona.msus.edu/is-f/library-f/webind2/webind2.htm, 1997.

[MHB+97] J. Moore, E. Han, D. Boley, M. Gini, R. Gross, K. Hastings, G. Karypis, V. Kumar, and B. Mobasher. Web page categorization and feature selection using association rule and principal component clustering. In *7th Workshop on Information Technologies and Systems*, Dec 1997.

[MS96] Y. S. Maarek and I.Z. Ben Shaul. Automatically organizing bookmarks per content. In *Proc. of 5th International World Wide Web Conference*, 1996.

[PE95] M. Perkowitz and O. Etzioni. Category translation: learning to understand information on the internet. In *Proc. 15th International Joint Conference on AI*, pages 930–936, Montral, Canada, 1995.

[Por80] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[SM83] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, 1983.

[TSM85] D.M. Titterington, A.F.M. Smith, and U.E. Makov. *Statistical Analysis of Finite Mixture Distributions.* John Wiley & Sons, 1985.

[WP97] Marilyn R. Wulfekuhler and William F. Punch. Finding salient features for personal Web page categories. In *Proc. of 6th International World Wide Web Conference*, April 1997.

[WVS+96] Ron Weiss, Bienvenido Velez, Mark A. Sheldon, Chanathip Nemprempre, Peter Szilagyi, Andrzej Duda, and David K. Gifford. Hypursuit: A hierarchical network search engine that exploits content-link hypertext clustering. In *Seventh ACM Conference on Hypertext*, March 1996.