

WordSieve: Learning task differentiating keywords automatically

Travis Bauer
Sandia National Laboratories
tlbauer@sandia.gov

August 1, 2003

This research¹ tests the effectiveness of using a competitive learning technique to automatically differentiate a user's task contexts during information access. Differentiation of a user's various task contexts can be used to provide customized information. The user provides implicit clues regarding task context by a tendency to access documents of similar topics together in groups. In such an environment, the frequency of task specific terms in the stream of documents accessed will change over time in a way that correlates to the changes in task context. The user does not have to do anything special to produce these term occurrence patterns, they are a byproduct of normal information access. Utilizing these implicit clues, a recommender system could extract task specific keywords. These keywords could be used to select potentially relevant documents to suggest to the user. To avoid overburdening the user, the keyword extraction process should be automated, requiring minimal user feedback, if any.

Many successful systems have been developed which try to extract terms for similar purposes. However, such systems typically use indexing techniques developed for static corpora where all documents are available to the indexing algorithm in advance, giving the algorithm a "bird's eye view" of the data. Such algorithms do not typically take advantage of the implicit clues in the ordered series of documents accessed by the user nor are they inherently well suited for adapting to a user's interests changing over time.

Our research has focused on the development of an unsupervised, competition based feature extraction algorithm called WordSieve which takes advantage of these implicit clues provided by the user. It learns to differentiate task contexts of users engaged in information access. It does not require user feedback, explicitly defined categories, or other external information regarding when contexts shift. It learns solely based on the stream of documents the user accesses. WordSieve learns task contexts by discovering terms with certain occurrence patterns which we believe correlate to context shifts. These terms are discovered via a competitive process where terms compete for limited space in a set of units. Terms which win this competition are used in term vectors to constitute user profiles and can be used to generate queries and indices.

¹This research was conducted under David Leake at Indiana University.

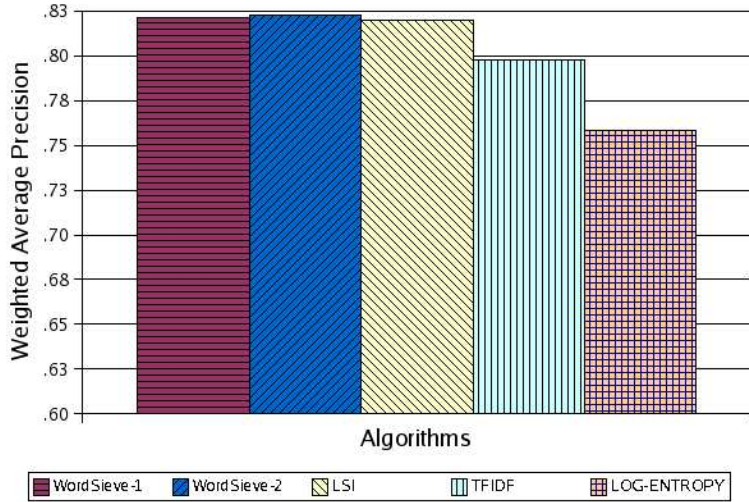


Figure 1: Performance with Browsing Data

WordSieve thus provides the ability to adapt to noisy, real world input without reliance on computationally intensive statistical summaries of large sets of data and explicitly defined categories.

Two attributes of WordSieve make it well suited for personal information retrieval. First, by using a competitive, stochastic process, WordSieve learns in real time using only local information. Thus, it does not need a static corpus to analyze. This helps it adapt to the changing interests of the user. Second, by taking advantage of the order in which a user accesses documents, it can effectively identify changing task contexts.

In order to test WordSieve, we collected a set of web browsing data. We have also used a set of newsgroup data. Both sets of data contain overlapping topics. Using this data, we tested the ability of WordSieve to index and retrieve documents according to the context within which the document originally appeared and compared the performance to that of other algorithms.

In previous literature [1, 2, 3], we have discussed the performance of initial implementations of WordSieve and its performance on some of the data mentioned above. This presentation will review the ideas behind WordSieve, discuss its latest implementation, and compare its performance to three common indexing techniques based on global information: Term Frequency/Inverse Document Frequency, Log-Entropy, and Latent Semantic Indexing. The new version of WordSieve outperforms the earlier version and reliably outperforms the other algorithms in our tests as well. Average weighted precision results on two precision/recall tests are shown in figures 1 and 2. This suggests that competitive learning techniques can be used to gather task context information implicitly provided by the user and use this information to improve information indexing and retrieval.

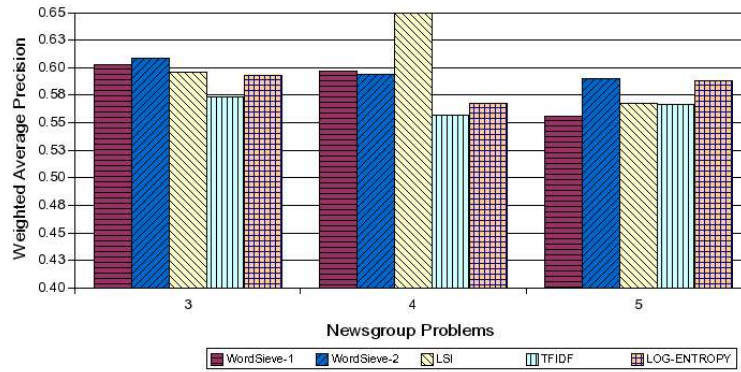


Figure 2: Performance with Usenet Data

References

- [1] T. Bauer and D. Leake. A research agent architecture for real time data collection and analysis. In *Proceedings of the Workshop on Infrastructure for Agents, MAS, and Scalable MAS*, 2001.
- [2] T. Bauer and D. Leake. Wordsieve: A method for real-time context extraction. In *Modeling and Using Context: Proceedings of the Third International and Interdisciplinary Conference, Context 2001*, pages 30–44. Springer-Verlag, 2001.
- [3] T. Bauer and D. Leake. Using document access sequences to recommend customized information. *IEEE Intelligent Systems*, 17(6):27–32, Nov/Dec 2002.