



Collection Management White Paper

Executive Summary

A significant portion of searching on the Web involves the creation by a user of a corpus of information about a narrow topic. The user often refines and adds to this corpus over an extended period of time, as more is learned. Current search models are very weak for this type of user, both in the discovery of new items to add to the corpus and in the use of that corpus by other, similarly-minded users. Yoogli's sense-matching technology enables users to quickly locate resources which match a corpus, both by topic and by significant phrase. This corpus is referred to as a collection, since it usually consists of a group of URLs. Yoogli also provides the ability for users to self-identify using their own collections, and matches other users (or simply their collections) to provide an instant gain in accumulated knowledge.

Application – Collection Manager

Yoogli's Collection Manager is designed for a community of users that assemble separate collections of data, each on a single topic. A big problem with the search models now being used to assemble these collections is that the context and content of the collection being built cannot be used to locate new resources that might belong - only key words can be used in the user's search query. No matter what the user is actually looking for, if a keyword that the user thinks will bring results appears on a completely unrelated page, that page will be returned anyway. The user is forced to manually classify pages by semantic domain before he can even begin the task of seeing if the page is related to the current collection of interest and useful to add to that collection.

Even when the collection is complete, it stays resident on the user's computer, being of no use to anyone else. It doesn't matter if multiple users are researching the exact same topic - all of the work needed to complete that research must be done by each user independently. For each user that successfully builds a collection, there are many users who would be interested in having the same collection. Unfortunately, those users lack either the ability or the discipline to complete the required research tasks.

Yoogli's semantic matching technology has enabled the creation of a Collection Manager application that possesses none of these flaws. Users can quickly locate URL resources that are semantically aligned with their current collection, and that collection can be effective in such a search even if it only contains one URL! Collections are stored and manipulated on a central server, so they can be made accessible (at the user's option) to other people that might be interested in the same topic. Since collections usually consist of multiple URLs that densely cover a specific portion of the semantic universe, they can be matched much more accurately than single URLs. As a result, it is possible to measure the semantic distance between two individual collections, as well as two sets of collections. This provides a measure of user similarity, allowing users to contact other people or make themselves available for such contacts from people who have created specific types of collections. This semantic partitioning of a social network is quite revolutionary when compared to the unfiltered (or attribute-filtered) social networking applications of today. Even if the social networking aspect is removed from the Collection Manager, it is still possible to derive value from anonymously constructed collections.

Collections can be returned as the result of a standard keyword query by linking the context of each returned URL with the most similar collections in the database.

Semantic Matching

To accomplish semantic matching between collections and URLs, it is necessary to provide a mathematical representation of context for each entity of interest. Any block of text, no matter how large, can be semantically classified using a numerical vector that specifies membership in each of some number N of domains, and some number M of significant phrases which occur in the text of the URL and which are semantically aligned with the overall text of the block. For this paper, a domain can be defined as a partition of some semantic space of interest, where the partition is defined using an enumeration of semantic entities that occur exclusively or primarily in that part of the overall space. A URL can be considered to belong strongly to that domain if it contains a significant number of the semantic entities that make up that domain. Conversely, a URL would be considered to belong weakly or not at all to a domain if it contained few or none of the semantic entities used to define that domain.

Similarly, a person can be assigned a membership vector to the same set of domains simply by examining their set of collections. Since every URL or even entire web site can be assigned a vector, assigning a person a similar vector is accomplished by simply combining the vectors from the collections they build in some way. This combination can be done in many different ways depending on the application, including (but not limited to) averaging, averaging with hysteresis, weighted combination, or noise-adding averaging. This vector created for a person is called a "personalization vector". Each collection, if the collections are semantically distinct, can be considered to be an "aspect" of the person. The person can use these aspects to assist in search operations, especially by removing irrelevant topics from the search results.

With a vector created for each collection of a user and a vector for every web page and site, implementation of a Collection Manager becomes nothing more than an augmented vector matching system. The vector of the collection of interest and a potential target's vector can be multiplied together using a dot product, a weighted dot product, or some other method, resulting in a match value M . Search results are ordered by M , meaning that the user will find much more relevant results on the first page of search results. Aligned key phrases stored with each entity can also be used to narrow the match when the vectors are sufficiently close.

Implementation

The Yoogli Collection Manager system starts with no user collections and zero knowledge about the user. When the user begins creating collections, a new record is created in the database for that user. Every time the user adds a new link to one of his collections, the contents of that page are used to update the collection and aspect of the user in the manner summarized above.

Users who log into the Collection Manager have the option of performing a normal keyword search, a collection-enhanced keyword search, or a collection only (no keyword) search. A collection-enhanced search pre-screens the universe of URLs before the keyword is applied, in such a way that only semantically related URLs are

considered for the search. A collection-only search performs the same operation, but then uses the semantically aligned key phrases of the originating collection as further filters in the result ordering. This is similar to performing hundreds of simultaneous keyword searches, but keeping only the most relevant results from each one. Users need not be logged in to realize the benefits of collection-enhanced search. While they cannot modify existing collections or add new ones, they still can get collections in their search results and traverse the community of collections by performing collection-enhanced and collection-only searches.

Operations

Implementation of a Collection Manager can take place in a URL universe of any size. From isolated corporate networks to the entire Web, the operation is the same. The only requirements are that the base domains be adequately defined, and that the complete set of available URLs be semantically indexed before search operations commence.

Conclusion

Yoogli's Collection Manager and related technology provides a powerful set of search tools to users who are doing topical research. Users no longer need restrict their results horizons to the first few pages of unclassified keyword results. They can use semantic prescreening and aligned key phrase matching to much more effectively search large databases and obtain results. Finally, all users can benefit from each other's efforts as collections become new targets of search results.