As a leading provider of enterprise information access solutions, Endeca has long been interested in automatic techniques for document annotation. Endeca’s approach emphasizes interactive and exploratory search, in the spirit of reference librarians working with information seekers to help elaborate and better understand their information needs. Since this interaction is data driven, augmenting the content through annotation dramatically enriches the dialog between the user and the system.

Many state-of-the-art techniques for document annotation are supervised, relying on models derived from manually created training data. However, enterprises are often unable or unwilling to incur the costs associated with such training. Instead, enterprises prefer unsupervised techniques that require minimal human involvement, are broadly applicable across domains, and produce sufficiently accurate results.

A natural starting point for such an unsupervised approach is terminology discovery, a technique that distills a vocabulary of terms from a collection of documents. This vocabulary then becomes the basis for annotating individual documents. The two-part approach is highly suitable for supporting exploratory search, particularly if the vocabulary is distilled to ensure that terms both are human-consumable and have high discriminatory power in the document collection. Such terms help users discover related documents, and, more generally, support the information seeking process.

Endeca has successfully applied terminology discovery to annotate a variety of domains. Applications use the annotations to improve relevance ranking, as well as to enable interactions that include progressive query refinement, similarity search, and clustering.

One of the problems with unsupervised terminology discovery, however, is that even the best linguistic and statistical algorithms are no substitute for domain knowledge. As a result, they may fail to discover useful terms, and may instead discover terms that, despite their statistical properties, are not valuable to human information seekers.

To address this concern, Endeca has developed a new approach that bootstraps on human-generated sources of terminology, such as user-supplied document tags and search logs, in combination with the collection itself. This approach extracts a vocabulary from these human-generated sources and then intelligently applies the vocabulary to the overall document collection.

We will demonstrate the success of this approach in the context of two very different domains: the ACM Digital Library and a leading provider of sports news coverage. We will also discuss recent work on applying document annotations to two distinct exploratory search problems: summarizing the current results, and providing the best directions for further exploration.

Note: this presentation represents joint work with Joyce Wang and Vladimir Zelevinsky.