Communications of the IIMA

Volume 10 | Issue 3 Article 3

2010

DEAR: A New Technique for Information Extraction and Context-Dependent Text Mining

Tod Sedbrook
University of Northern Colorado, Monfort College of Business

Jay M. Lightfoot
University of Northern Colorado, Monfort College of Business

DEAR: A New Technique for Information Extraction and Context-Dependent Text Mining

Tod Sedbrook
University of Northern Colorado, Monfort College of Business
tod.sedbrook@unco.edu

Jay M. Lightfoot
University of Northern Colorado, Monfort College of Business
jay.lightfoot@unco.edu

ABSTRACT

The desire to store and the need to use electronic data has greatly increased as the power, availability, and connectivity of computers has grown. A large portion of this data is in the form of unstructured text documents. Locating specific information within this amorphous mass of documents is an area of active research. Our contribution to this pursuit is the development of the Document Entity and Resolution (DEAR) system. This system combines semantic similarity matching as provided by the open source WordNet database with the ability to recognize named entities through the OpenCalais system. When used in concert, this provides a novel way for users to quickly find relevant content and detect and identify uniquely named entities within that content. The theory behind the system is defined and the working system is described. This system is then applied to a collection of assessment documents as a proof-of-concept test of its viability. The results are promising and indicate that further research is warranted.

INTRODUCTION

The availability and increased power of computer technology, the decreased cost of disk storage space, and the connectivity provided by the Internet has created a situation where vast amounts of unstructured text-based documents are being stored electronically. To access text-based information, data mining and text mining techniques are used. Data mining techniques require data to be stored in well organized, structured formats; however, text mining techniques are able to extract useful information from unstructured document collections. Because of this, text mining techniques are useful in processing these documents. The goal of text mining is to structure document collections to improve the ability of users to retrieve and apply knowledge implicitly contained within those collections (Ikonomakis, Kotsiantis, & Tampakas, 2005). Text mining proceeds through three phases to accomplish this goal: pre-processing, pattern discovery and visualization.

Within document collections, natural language signifies meaning within a maze of synonyms and domain specific terms (Blake & Pratt, 2001). The text mining pre-processing phase cleans and analyzes document collections to transform implicit meaning into normalized and explicitly structured concepts. Pre-processing challenges include defining ways to manage the

heterogeneity of terms and phrases that result from the increasingly dynamic and geographically dispersed contributions to document collections (Kwak & Yong, 2010).

The pattern discovery phase analyzes and derives distributions of concepts across document collections to help users filter and identify relevant documents. A challenge of pattern discovery is to retrieve manageable subsets of relevant documents and then alert users to further context dependent queries that may refine their initial results. A challenge of visualization, the third phase of text mining, is to support the user by providing dynamic graphs for visualizing relevant relationships among identified documents (Feldman & Sanger, 2009).

Taken together, these challenges provide the justification for this research project. Our goal is to design and explore a text mining system to quickly and easily retrieve data that has been stored in numerous text documents. Further, retrieval of this information should not rely on the end-user knowing all the various search terms under which it may have been stored. Rather, it should allow useful information to be located via search terms familiar to an end-user. The technique developed by this research will utilize two externally developed, freely-available, word context systems along with a custom search program written in F#, Microsoft's new declarative .Net language. The methodology is explained and demonstrated through a system designed to improve the semantic match between a document collection and an end-users knowledge requirements. This project designs and demonstrates techniques to bridge the gap between unstructured document collections and text-mining tools to support retrieval and evaluation of knowledge within those collections.

PROBLEM DESCRIPTION AND RELATED RESEARCH

In its simplest form, the problem addressed by this research is to develop a method that can find information stored in unstructured text documents using a query vocabulary which may not match that used in the stored documents. To illustrate this situation, assume that a new assessment office is created within a university. The purpose of the position is to coordinate the various programs offered by the colleges within the university and devise consistent assessment schemes that can be uniformly applied across all colleges. The initial task that must be performed by the new Vice-President of University Assessment is to determine how the existing assessment systems work. This requires searching through voluminous documents written by different authors in different disciplines over an extended period of time. Assuming that the documents are all available in a machine readable format, the first problem that is encountered is terminology. The simple query of, "What topics are covered in the syllabi for classes?" runs into the problem that within one college of the University, the periodic meeting of student and teacher for the purpose of instruction is referred to as a "course" whereas others refer to the same concept as a "class" or a "session." Similarly, a "topic" in one program could have the same meaning as "subject", or "theme", or "discipline" in another. Without some mechanism to deal with different vocabulary terms that refer to the same concept, the end-user is forced to repeatedly guess terms that may be applicable for keyword searches. Besides the obvious inefficiency of this method, there is also the very real probability that simple keyword searches will fail because the correct term was not guessed.

Researchers have devoted a significant amount of effort toward the goal of solving this problem. This research is generally listed under the categories of *text mining* or *information extraction*. Within this field of study are some concepts that require explanation. The remainder of this section will address these topics.

Basic Information Extraction Concepts

Information extraction (IE) is concerned with extracting relevant fragments of information from unstructured text documents in order that the fragments be automatically processed further, for instance, to answer user queries (Kauchak, Smarr, & Elkan, 2002). There are three basic approaches to information extraction: rule-based, statistical methods, and knowledge-based methods (Siefkes & Siniakov, 2002). Rule-based approaches use template patterns to analyze text structure in an attempt to find and interpret the relevant information. These patterns can be hand entered or machine generated. The statistical methods use mathematical calculations to predict the likelihood that the desired information is present based on the surrounding text and context predictors. The knowledge-based approaches utilize external knowledge to help categorize and classify the text so that relevant information can more easily be found (Siefkes & Siniakov, 2005). The external knowledge can be machine generated or acquired from an existing lexical ontology. An ontology is "an explicit, formal specification of a conceptualization that is shared among a group" (Gruber, 1993). A lexical ontology would be one that provides a shared vocabulary for understanding the language domain. Of the three, the knowledge-based approaches are most applicable to our research because they deal more directly with information *retrieval* rather than the more elaborate information *understanding* to which the other two methods aspire.

Regardless of the method selected, all information extraction systems must deal with the problem of ambiguity: the existence of multiple word sense meanings within the context of the text (Cohen & Hunter, 2008). Many techniques have been developed to deal with ambiguity; two of the most common are expansion and partitioning. Expansion (also known as *enhancement*) can be applied to queries and the source document set. The processes in both cases are similar in that the set of searchable terms is increased through the use of synonyms, hypernyms, and hyponyms. A *hypernym* is a term that is semantically more generic than another term (e.g., "meat" is a hypernym of "beef"). A *hyponym* is a term that is semantically more specific than another term (e.g., "apple" is a hyponym of "fruit"). Taken together, one can create a generalization/specialization hierarchy of terms with semantically related meanings. The intent of the expansion is to include semantically similar terms and phrases into either the query question or the source documents so that it is no longer required that the person asking the question use the exact same vocabulary as the document creator. This increases the likelihood of a match and decreases the impact of ambiguity (Hotho, Staab, & Stumme, 2003).

The inverse of expansion is text partitioning (also known as *categorization*). This process is also useful in IE systems because it deals with ambiguity in a different way. In this process, terms are grouped into categories based upon the semantics of the context. In effect, the number of terms decreases as more generic concepts subsume them. This can be done in either a supervised or an unsupervised mode. *Supervised* partitioning implies manual clustering; hence, it requires significant human effort to achieve. *Unsupervised* partitioning is the computer-driven, automatic

mode that is best suited for large document sets. The unsupervised mode also normally employs either algorithmic techniques based on statistical analysis of the source text or background knowledge from an existing lexical ontology. As would be expected, there are many variations and approaches to implementing the expansion and partitioning techniques (Amine, Elberrichi, & Simonet, 2009).

Text Preparation

Before any of the information extraction methods are implemented, the target text is normally prepared through a series of steps designed to remove semantic "noise." In the context of text mining, these are terms and phrases that do not add any semantic content. For the knowledge-based approaches, this begins by removing punctuation, numbers, and converting all text to lowercase. Following this, common preparation steps include (Hotho et al., 2003):

- Stopword removal – Stopwords are common words and phrases that do not add any semantic information to the text. For example: *a*, *an*, *the*, and *is* are all stopwords. In addition, commonly used domain-specific words can also be stopwords. For example: company names, email addresses, and proper nouns also could be removed.

- Stemming and lemmatization – Together, these two processes are said to *normalize* the text. Stemming is the process of transforming terms into their root forms (Sahami, 1999). For example: the stem of *running*, *runner*, and *ran* is **run**. Lemmatization is the process of transforming words into a standard format where nouns are singular and verbs are in the infinitive form (Airio, 2006). Of the two operations, stemming is the more common because lemmatization is more complex and requires more knowledge about the context of use.

- Pruning – This step removes rare terms that do not add significant semantic content. There are various methods to determine the definition of "rare." In some cases, this step is combined with text partitioning to prune via categorization (Liu, Liu, Yu, & Meng, 2004). The method chosen can dramatically affect the results of the information extraction activity. This is an area of active research.

- Weighting – Terms that remain after pruning must be weighted to determine their importance within the context of the document or query. As with pruning, there are multiple techniques and algorithms used for this process and the method chosen greatly impacts the results (Amine et al., 2009; Hotho et al., 2003; Baeza-Yates & Ribeiro-Neto, 1999). Freely available domain ontologies exist that provide weighting values based on standard English usage, so it is not necessary to generate your own (Miller, 1990; Fellbaum, 1998).

Text Representation and Analysis

Following the preparation steps, information extraction requires that the normalized terms be represented in a format conducive to analysis. Multiple representations have been devised; however, two of the most common are the "bag of words" vector model and the ontology-based

representation. Generally speaking, the "bag of words" model treats the text as a vector where each component is a term from the source text. These vectors also may hold other information such as the frequency of occurrence of the term and the weighting of the term in the document set (Amine et al., 2009). In this representation, text expansion is achieved by adding new elements to the vector. Text partitioning can be implemented by replacing terms with category components. Determining terms common to multiple documents is easily performed with an intersect function while merging all terms in multiple documents is possible with a union operation. This representation is extremely flexible and intuitive.

The ontology-based representation also uses a vector for the terms, frequencies, and weights. In addition, this model adds an ontology structure to provide *concept* (i.e., category) information. This additional information provides two key benefits: 1) it resolves synonyms between term categories, and 2) it allows a hierarchy of more general and more specific terms (i.e., hypernyms and hyponyms) to be available (Hotho et al., 2003). This additional information helps resolve ambiguous semantic relationships between words which, in turn, improve the ability of the system to deal with queries involving inconsistent vocabulary. Sophisticated natural language processing systems generate the ontology model directly from the document set; however, this is time-consuming and overly complex. An alternative approach that is both popular and effective is to utilize existing general-purpose lexical ontologies. Freely-available lexical ontologies such as WordNet (Miller, 1990; Fellbaum, 1998), HowNet (Dong, 1999), and SENSUS (Knight & Luk, 1994) have been successfully used in text mining research. It has been shown through this research that the addition of the background knowledge provided by an ontology improves the overall semantic matching performance of a system (Hotho et al., 2003).

Analysis of the text representation proceeds by performing various expansion, pruning, matching, partitioning, and disambiguation operations. Expansion is useful for finding synonyms and, when used in concert with partitioning, can determine conceptual similarity between disparate terms and phrases. Pruning can remove terms from the vector space that are judged to add no semantic content while matching helps find related concept sets. Partitioning is used to categorize terms into more general hypernyms or more specific hyponyms so that the lexical knowledge stored in the ontology can be utilized.

The final task of disambiguation is generally intended to remove lexical "noise" that may have been added by partitioning a large number of terms into a single category. While partitioning is useful and necessary, it often adds words that are out-of-step with the semantic context of the source text. Disambiguation selects the most appropriate terms and increases their value weights. Inappropriate terms are also identified and removed from the vector space (Hotho et al., 2003; Ide & Veronis, 1998). More complex schemes deal with phrases and context analysis. To aid in the disambiguation process, some ontologies (e.g., WordNet) have pre-built domains of tagged categories. These help guide the word sense disambiguation process (Kolte & Bhirud, 2008; Bentivogli, Forner, Magnini, & Pianta, 2004). They can also be used as an addition layer of processing to improve the semantic performance of the information extraction system.

The following sections will describe the specific information extraction system that was created by this research project. This system has many of the capabilities described above and, we believe, is unique in its novel approach to information extraction.

DESCRIPTION OF THE TECHNIQUE

Document Ranking Algorithms

Named entity recognition (NER) classifies elements in unstructured text into semantic metadata by recognizing the names of persons, organizations, employment position, industry terms, monetary values etc. We apply OpenCalais, a web service provided by Clear-Forest, a Thomson Reuters Company, to recognize named entity features in a set of documents. Let $\overrightarrow{D} = \{d_1, d_2, \ldots.., d_k\}$ be a set of k documents and consider the OpenCalais web service as providing a characteristic function allowing mapping of a document's word phrases (p) to its corresponding named entities (e): $\overrightarrow{E}k = \{(p_1,e_1), (p_2,e_2),,\ldots.., (p_p,e_p), \}$.

Further analysis applies $\overrightarrow{D}$ to define A* representing a matrix of a document's unique words (w_{kn}) along with each word's numeric frequency (k_{kn}) in each document in $\overrightarrow{D}$. Within the domain A* the following operations are performed for each word represented in lower case in A*. First, let $\overrightarrow{Z} = \{c_1, c_2, \ldots.., c_m\}$be a set of common words in the English language and the vector subtraction c represents c: $(A^* - Z^T) \rightarrow A^{**}$ where:

$$A^{**} = \begin{pmatrix} \{w11,F11\} & \cdots & \{wk1,Fk1\} \\ \vdots & \ddots & \vdots \\ \{w1k,F1k\} & \cdots & \{wkn,Fkn\} \end{pmatrix} \quad (1)$$

For each word in A** we then apply WordNet to identify nouns and for each noun we form three concept vectors corresponding to that noun's synonyms, $\overrightarrow{Sy}_{kn} = \{s_1,s_2,\ldots\ldots s_3\}$, hyponyms, $\overrightarrow{Ho}_{kn} = \{ho_1,ho_2,\ldots\ldots ho_3\}$, and hypernyms $\overrightarrow{He}_{kn} = \{he_1,he_2,\ldots\ldots he_3\}$. Synonyms relate directly to the meaning of each word, hyponyms capture more specific meaning, and hypernyms represent more general meaning of terms. The above forms the representation matrix **X** where for each document (k)

$$\mathbf{X_k} = \{ \overrightarrow{E}_k, \begin{pmatrix} \{wk1,Fk1\}, \overrightarrow{Sy}_{k1,} \overrightarrow{Ho}_{k1,} \overrightarrow{He}_{k1}\} \\ .. \\ \{wkn,Fkn\}, \overrightarrow{Sy}_{kn,} \overrightarrow{Ho}_{kn,} \overrightarrow{He}_{kn}\} \end{pmatrix} \}$$

(2)

The user defines a vector of query terms $\overrightarrow{Q} = \{q_1, q_2,\ldots..,q_i\}$ and we define a mapping (ε) into the document space $\overrightarrow{D}$ to rank to identify each document' s relevancy to the query:

$$\varepsilon: \{ \overrightarrow{Q}, \mathbf{X} \} \rightarrow \{d_j,f(j)|\forall_j \} \quad (3)$$

where f(j) represents the result of a confidence function, returning real values [0,1], representing the relevancy of the query mapping $\overrightarrow{Q}$ for each document in $\overrightarrow{D}$.

The mapping calculates the occurrences (Occ) as the sum of the frequency of a document's (j) matching words (m) by applying the document's representation matrix:

$$Occ(\overrightarrow{Q}, \mathbf{X}_j) = \sum_{q=1}^{q=i} \left(\sum_{m=0}^{m=n} \text{Fjm} * I(w\text{jm})\right)$$

(4)

$$\text{where } I(w\text{jm}) = \begin{cases} 1 & if\ q_i = w_{jm} \cup \\ & q_i \in \overrightarrow{Sy_{jm}} \cup \\ & q_i \in \overrightarrow{Ho_{jm}} \cup \\ & q_i \in \overrightarrow{He_{jm}} \cup \\ & q_i \in \overrightarrow{E_j}(p) \\ 0 & otherwise \end{cases}$$

The identity function $I(w\text{jm})$ determines when each word within a document is included in the (Occ) sum. That is, for a document (j) the frequency of each of its (m) words was included in the sum when a query term (q_i):

1. identically matched to a document's word or
2. was a member of the synonyms $\overrightarrow{Sy}_{\text{jm}}$ set defined for that word or
3. was a member of hyponym ($\overrightarrow{Ho}_{\text{jm}}$) set defined for that word or
4. was a member of the hypernym $\overrightarrow{He}_{\text{jm}}$ set
5. was a member of a recognized concept phrases (p) within ($\overrightarrow{E}_{\text{j}}$).

The relevance ranking of the document (j) to the query vector is then

$$f(j) = \frac{Occ(\overrightarrow{Q}, \mathbf{X}_j)}{\sum_{d=1}^{d=k} Occ((\overrightarrow{Q}, \mathbf{X}_d)} \qquad (5)$$

Documents are ranked according to $f(j)$ and a sorted document list is displayed to the users. The user can then further refine the difference between each ranked documents' name entities. By selecting a document (d_j), the system performs a set difference operation to determine and display the uniquely appearing OpenCalais named entities in the selected document:

$$\overrightarrow{E}\text{j(unique)} \;=\; \overrightarrow{E}\text{j} - \sum_{\text{d=1}}^{\text{d=k}} \overrightarrow{E}d_{\ (\text{d}\neq\text{j})} \qquad (6)$$

System Architecture

The search algorithms described above were implemented within the **D**ocument **E**ntity **A**nd **R**esolution (DEAR) system. The system consists of several modules and Figure 1 displays the overall architecture. The first-tier of the architecture consists of the Query and Analysis user interfaces. The Analysis interface allows users to specify the locations of documents on either the local file system or on the web and then interacts with the Analysis manager and remaining

modules to analyze and create a representation matrix (equation 2) that characterize each document within a collection. The Query interface first allows the user to define a set of query terms that characterize their interests. The user may also explore the WordNet categories to refine the meaning of their query terms. For example, the user may propose the query term "product" and WordNet will propose more general search terms (good, outcome), synonymous terms (production, ware) or more specific terms (freight, product line). After entering their selected query terms, the system's Query manager presents ranking scores and graphics representing the relevance of matching documents. The user can then select a returned document and the Query manager computes either the unique named entities for that document or all recognized entities in that document.

The Analysis manager consists of multiple routines that support the implementation of equations 1 and 2 defined above. The Analysis manager updates a relational database system that manages relations between documents, name entity categories, a document's word frequencies, and each word's expanded WordNet synonyms, hyponyms and hypernyms.

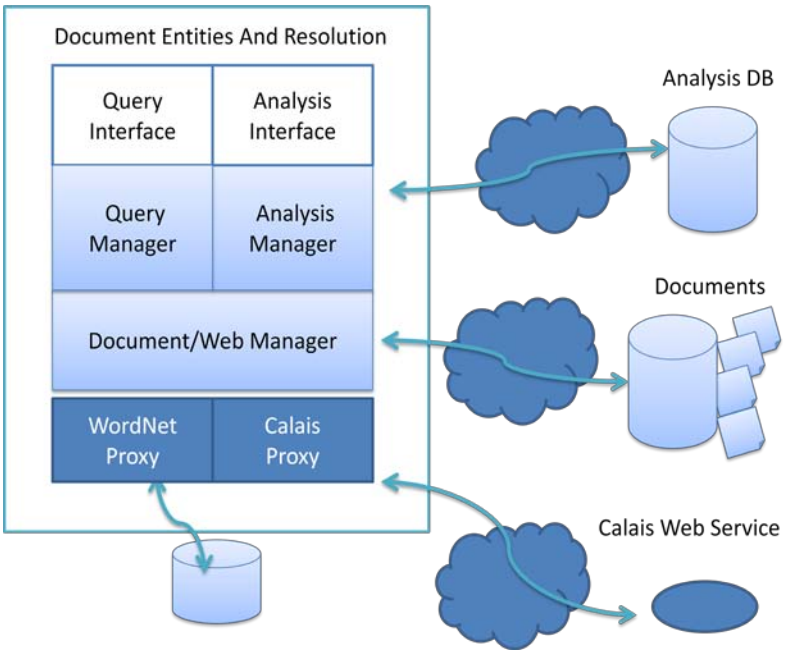

Figure 1: DEAR Architecture and Related Resources.

The Query manager is responsible for providing an implementation of equations 3 through 6. The Query manager applies the user queries to form sorted rankings of matching documents and presents the set of unique named entities within a selected document.

The Document and Web manager are responsible for retrieving the documents from the web and providing the Query and Analysis manager with transparent access to the WordNet and OpenCalais proxies. The WordNet proxy supports access to the WordNet data store for forming and matching synonyms, hyponyms and hypernyms. The OpenCalais proxy presents an interface to a web service that identifies sets of named entities occurring within the documents.

The architectural implementation takes advantage of the latest .Net technologies including F#, Windows Presentation Foundation, and the .Net 4.0 entity data model. In addition, several open-source libraries provided the runtime proxies to access WordNet and the OpenCalais web service. The following describes an example of the system's use in a proof-of-concept application.

System Application

The DEAR system analyzed a collection of fifteen documents containing minutes of a College's assessment committee meetings. The documents were available in Microsoft Word format where each was first converted into plain text and made available at a location on the local file system. The DEAR system identified 212 OpenCalais recognized name entities, and a total of 34,581 expanded synonyms, hyponyms and hypernyms resulting from 3,198 words within the document collection.

Figure 2 demonstrates the use of WordNet categories to assist a user in refining their query. Consider a user that is interested in assessment minutes that discussed ways to measure a student's moral behavior. The system first suggests other synonyms, hyponyms and hypernyms that may better express a user's intentions. For example, "morals" in the user query triggers suggestions for closely related synonyms such as "ethics." The user may accept or reject the system's suggestions.

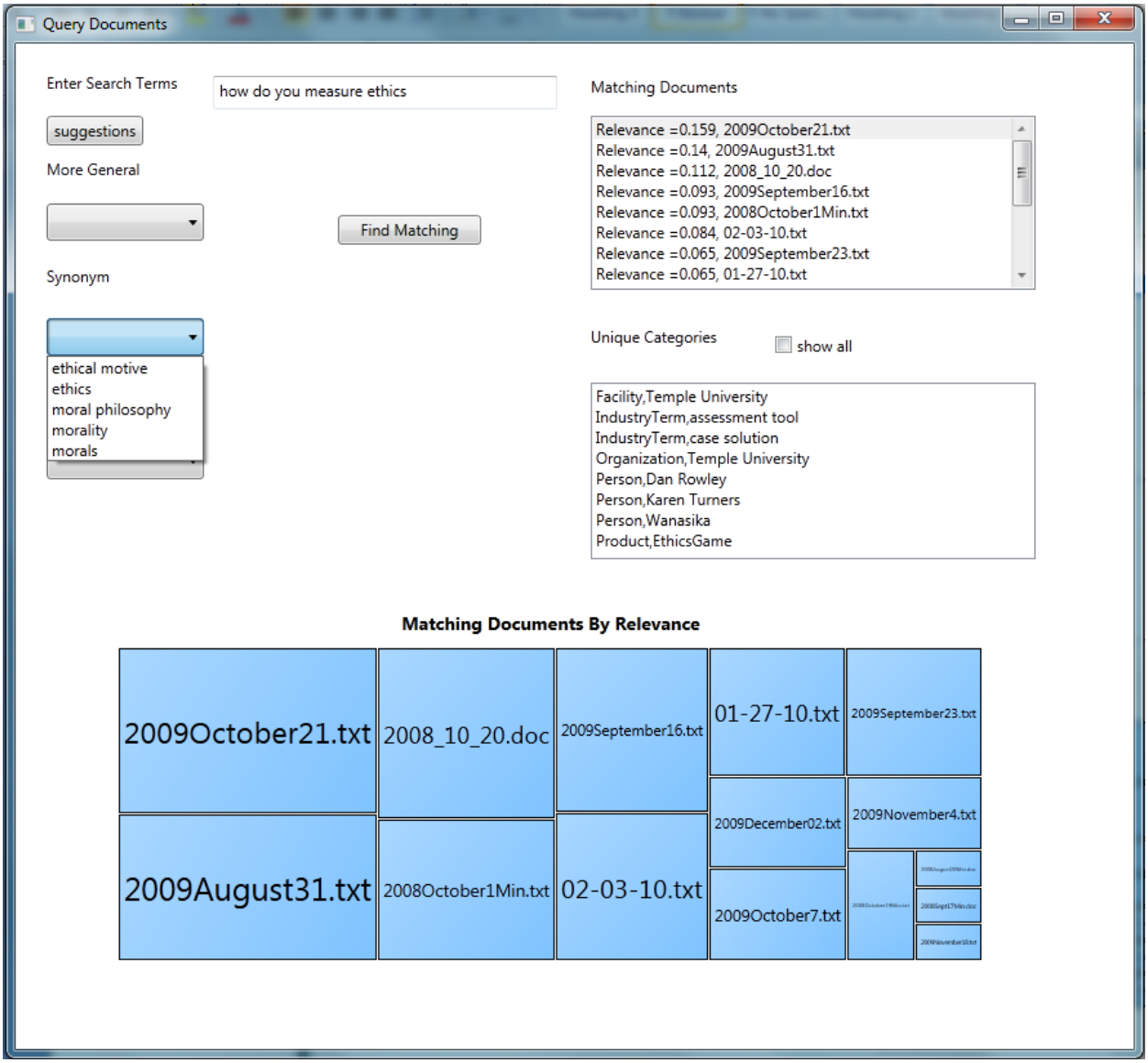

Figure 2: Results of Search for Query Terms: how do you measure student ethics.

Figure 2 shows the result of the query "how do you measure student ethics?" The system ranked the documents according to its relevance ranking (equation 6) and reported that the October 21st minutes had the best match. Each matching document's relevancy ranking is presented both numerically and visually through a *treemap* graphic. The treemap displays a tiling of nested rectangles, where the size of each labeled rectangle is proportional to that document's relevancy to a query.

Users selecting the October 21st minutes are presented with the set of named entities that appear only in the selected document. The system recognized, for example, that the October 21st minutes uniquely contained the industry terms "assessment tool" and "case solution", and the product "EthicsGame." The August 31st minutes, while close in numerical query relevance to the October 31st minutes, contained entities relating to "Updated Goals and Objectives." A user can select and quickly evaluate the set of differential entities in the matched document set and zero-

in on the document that best matches their needs. In one case, a user may be interested in tools and techniques to measure a student's ethical understanding and first investigate the October 21st minutes. In another case, a user may be interested in what ethical assessment goals and objectives are measured and first investigate the August 31st minutes.

Figure 3 displays additional results from the document set for a query concerning financial sources. This time three documents were ranked equal in relevance: November 4, December 2, and January 27 minutes. However, the November 4 committee minute's document uniquely recognized the position "Dean for Funding." In addition, Figure 3 displays potentially more specific query terms that the users may consider in the place of "finance."

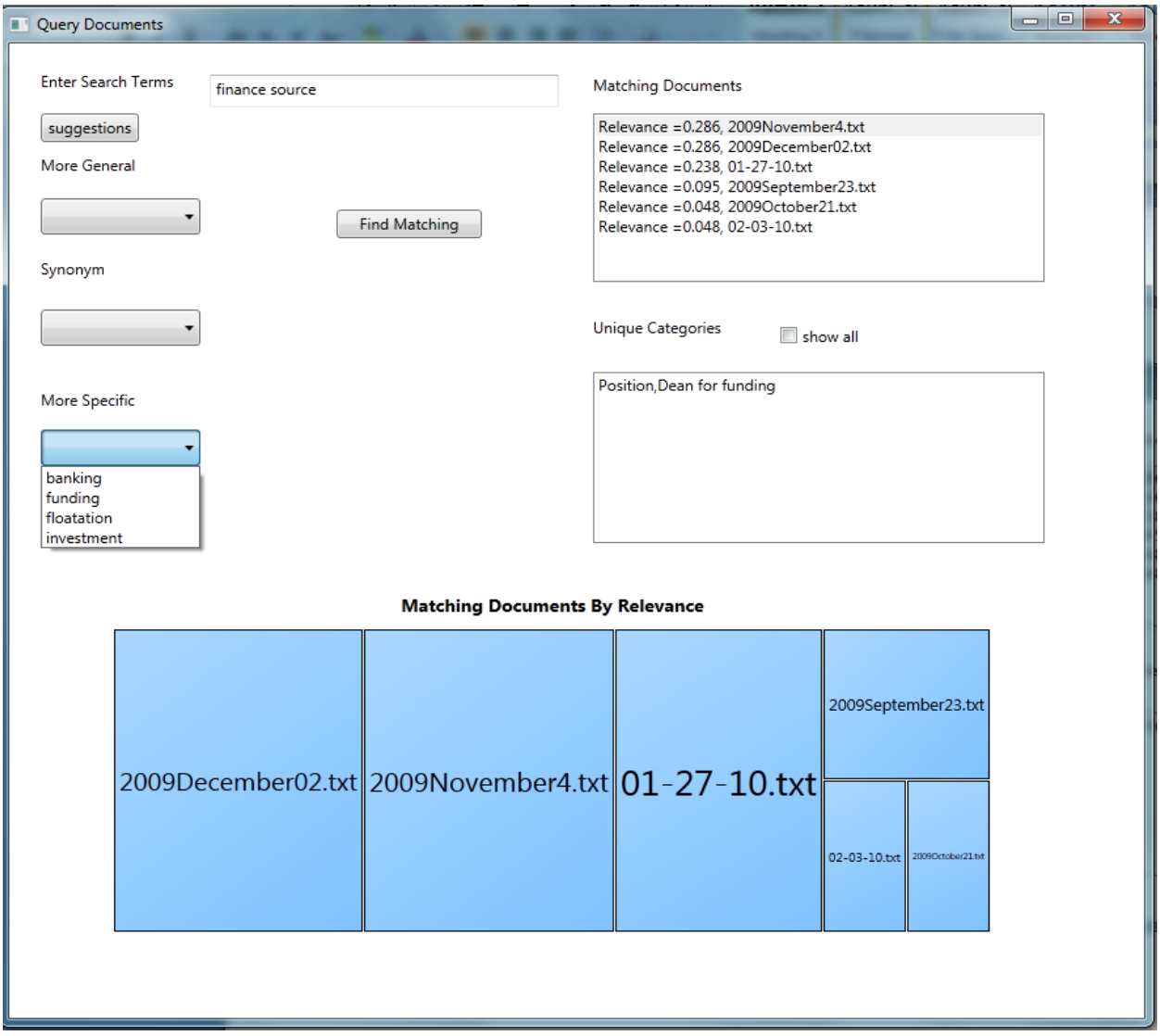

Figure 3: Search for Keywords: finance source.

DISCUSSION AND FUTURE WORK

Eighty percent of the time two users disagree on a common term to describe an object or concept (Furnas, Landauer, Gomez, & Dumais, 1987). A user's initially proposed query terms have only a moderate chance of directly matching a relevant document terms, but expanding the search space to include multiple meanings improves search success (Nikolova, Ma, Tremaine, & Cook, 2010). Search and understanding can also be improved by providing the user with additional background information including semantic annotations with named entities (Nauerz, Bakalov, König-Ries, & Welsch, 2008).

Our goal was to improve the user experience for searching specialized document collections. The DEAR system's architectural design goal was to blend together the open source projects WordNet and OpenCalais to provide a novel way for users to quickly find relevant content and to identify unique named entities within that content. Other techniques require first furtively expanding a user's query terms through WordNet and then applying the expanded query terms to search a document collection (Gong, Muyeba, & Guo, 2010). In contrast, we present the user with possible hypernyms, synonyms and hyponyms that relate to their search terms and allow them the opportunity to directly select new terms that may better express their intent. We, in addition, apply WordNet to preprocess and intelligently expand each noun within the document collection prior to the query to better represent the range of semantic meaning within documents. The preprocessing stage allows the system to quickly match the user's search terms to identify relevant documents from WorldNet's semantically preprocessed term sets.

OpenCalais represents a complementary way to increase the semantic content of the user's search through recognition of nearly 100 categories of named entities that could appear within a document collection. The user selects a document and the system identifies sets of semantic metadata representing named entities including people, organizations, and industry terms, that uniquely characterize that document and which do not appear within any other document in the collection. The system's semantic similarly matching facilities provided by WordNet, combined with OpenCalais's detection of name entity provides the user with powerful query tools to quickly identify relevant documents that match their needs.

The system is, however, not without limitations. Preprocessing and WordNet expansion of each document generates nearly 10 expanded terms for each noun appearing within the documents; consequently, the current system is limited in its ability to scale up for preprocessing collections containing thousands of documents. Future research is needed to provide better indexing and tuning to control WordNet's hyponym, synonym, and hyponym expansion in the document preprocessing stage. The preprocessing of documents through WordNet is relatively slow and there is a need to explore parallel algorithms to improve analysis speed in the document preprocessing stage.

Improvements to the user interface are also needed. We are exploring improvements to present snapshots of relevant sentences within matching documents. The snapshots will be ranked according the proximity of query words within sentences to allow the user to better judge the semantic content of documents matching query terms.

The construction and testing of the DEAR system is part of a larger project to improve the automated creation of domain ontology. Ontologies are often large and complex structures whose development and maintenance are difficult. The WordNet expansion and OpenCalais identification provides a rich semantic categorization of a document collection. Further research is needed to analyze resulting categories to distill and represent that knowledge in a formal and reusable ontology.

CONCLUSION

A vast amount of information is hidden within electronic data stored in the form of unstructured text documents. Accessing knowledge within unstructured document collections requires interface tools to help users initially define their query, logical tools to semantically match document concepts to a query request, and reporting and analysis tools that allow users to assess the relevancy of the resulting document matches. We developed the DEAR system to semantically enhance knowledge search and retrieval by combining the open source WordNet lexical database of English and the OpenCalais semantic metadata web service.

First, DEAR allows the user to refine their queries to better express their intentions with the assistance of WordNet's semantic categories. Second, each noun occurring within the document collection is preprocessed via WordNet to expand its meaning into related sets of synonymous concepts, more general concepts, and more specific concepts. The preprocessed set of expanded terms supports high fidelity matches between the concepts expressed within the user's query and the concepts contained in the document collection. Third, the OpenCalais metadata helps users recognize a document's name entities and investigate the semantic differences among documents that match their query's request.

The DEAR project is currently in the early stages of development. While a successful proof-of-concept system has been built and demonstrated, there are still scalability and performance issues that must be addressed. Future versions of the system will address these concerns and will move the use of the system toward the more ambitious goal of providing for the automated creation of domain ontologies.

REFERENCES

Airio, E. (2006). Word normalization and decompounding in mono- and bilingual IR. *Information Retrieval*, 9(3), 249-271.

Amine, A., Elberrichi, Z., & Simonet, M. (2009, March). WordNet-based text clustering methods: Evaluation and comparative study. *International Review on Computers and Software*, 4(2), 220-228.

Baeza-Yates, R.A., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. Harlow, UK; Pearson Education Limited.

Bentivogli, L., Forner, P., Magnini, B., & Pianta, E. (2004). Revising WordNet domains hierarchy semantics, coverage, and balancing. *Proceedings of the 20th International*

Conference on Computational Linguistics: Workshop on Multilingual Linguistic Resources. Geneva, Switzerland, 101-108. Retrieved from http://wndomains.fbk.eu/publications/Coling-04-ws-WDH.pdf.

Blake, C., & Pratt, W. (2001). Better rules, few features: A semantic approach to selecting features from text. In N. Cercone, T. Y. Lin, & X. Wu (Eds.), *Proceedings of the 2001 IEEE international Conference on Data Mining* (59-66). Washington, DC.

Cohen, D. B., & Hunter, L. (2008). Getting started in text mining. *PLoS Computational Biology, 4*(1), e20. doi:10.1371/journal.pcbi.0040020. Retrieved from http://www.ploscompbiol.org/article/info%3Adoi%2 F10.1371%2Fjournal.pcbi.0040020.

Dong, Z. (1999). Bigger context and better understanding: Expectation on future MT technology. *Proceedings of the International Conference on Machine Translation & Computer Language Information Processing*, Beijing, China, 17-25. Retrieved from http://www.keenage.com/papers/MTFuturetech.doc.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, MA; MIT Press.

Feldman, R., & Sanger, J. (2009). *The text mining handbook.* Cambridge, UK; Cambridge University Press.

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1987). The vocabulary problem in human-system communication: An analysis and a solution. *Communications of the ACM*, 30(11), 964-971.

Gong, Z., Muyeba, M., & Guo, J. (2010). Business information query expansion through semantic network. *Enterprise Information Systems*, 4(1), 1-22.

Gruber, T. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2), 199-220.

Hotho, A., Staab, S., & Stumme, G. (2003). Wordnet improves text document clustering. Proceedings *SIGIR Semantic Web Workshop*, Toronto, Canada. Retrieved from http://reference.kfupm.edu.sa/content/w/o/wordnet_improves_text_document_clusterin_97545.pdf.

Ide, N., & Veronis, J. (1998). Word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1), 2-40.

Ikonomakis, M., Kotsiantis, S., & Tampakas V. (2005). Text classification using machine learning techniques. *WSEAS Transactions on Computers*, 8(4), 966–974.

Kauchak, D., Smarr, J., & Elkan, C. (2002). Sources of success for information extraction methods. Technical Report CS2002-0696. San Diego, CA; University of California, San Diego. Retrieved from http://cseweb.ucsd.edu/users/elkan/BWI.pdf.

Knight, K., & Luk, S. K. (1994). Building a large knowledge base for machine translation. *Proceedings of the 12th National Conference on Artificial Intelligence, 1.* Seattle, WA, 773-778.

Kolte, S. G., & Bhirud, S. G. (2008). Word sense disambiguation using WordNet domains. *Proceedings of the First International Conference on Emerging Trends in Engineering and Technology,* Nagpur, Maharashtra, India, 1187-1191.

Kwak, J., & Yong, H. (2010). Ontology matching based on hypernym, hyponym, holonym and meronym sets in WordNet. *International Journal of Web & Semantic Technology* (IJWesT), 1(2).

Liu, S., Liu, F., Yu, C., & Meng, W. (2004). An effective approach to document retrieval via utilizing WordNet and recognizing phrases. *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* Yorkshire, UK, 266-272.

Miller, G. A. (1990). WORDNET: An online lexical database. *International Journal of Lexicography*, 3(4), 235-312.

Nauerz, A., Bakalov, F., König-Ries, B., & Welsch, M. (2008). Personalized recommendation of related content based on automatic metadata extraction. *Proceedings of the 2008 Conference of the Center for Advanced Studies on Collaborative Research: Meeting of Minds*, 57-71.

Nikolova, S., Ma, X., Tremaine, M., & Cook, P. (2010). Vocabulary navigation made easier. *Proceeding of the 14th international Conference on intelligent User interfaces*, Hong Kong, China, 361-364.

Sahami**,** M. (1999). Using machine learning to improve information access**.** Ph.d. thesis, Dept. Computer Science, Stanford University.

Siefkes, C., & Siniakov, P. (2005). An overview and classification of adaptive approaches to information extraction. *Journal on Data Semantics*, 4, 172-212.

This Page Left Intentionally Blank