Semantics-Empowered Text Exploration for Knowledge Discovery

Delroy Cameron
Kno.e.sis Center
Wright State University
3640 Colonel Glenn Highway
Dayton, OH 45435 USA
delroy@knoesis.org

Pablo N. Mendes
Kno.e.sis Center
Wright State University
3640 Colonel Glenn Highway
Dayton, OH 45435 USA
pablo@knoesis.org

Amit P. Sheth
Kno.e.sis Center
Wright State University
3640 Colonel Glenn Highway
Dayton, OH 45435 USA
amit@knoesis.org

Victor Chan
Air Force Research Lab
Wright-Patterson AFB
Dayton, OH 45433-5707 USA
victor.chan@wpafb.af.mil

ABSTRACT

The interaction paradigm offered by most contemporary Web Information Systems is a *search-and-sift* paradigm in which users manually seek information using hyperlinked documents. This paradigm is derived from a document-centric model that gives users minimal support for scanning through high volumes of text. We present a novel information exploration paradigm based on a data-centric view of corpora, along with a prototype implementation that demonstrates the value in content-driven navigation. We leverage semantic metadata to link data in documents by exploiting named relationships between entities. We also present utilities for gathering user generated navigation trails, critical for knowledge discovery. We discuss the impact of our approach in the context of knowledge exploration.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval Models; I.2.4 [**Knowledge Representation Formalisms and Methods**]: Semantic Networks—*Ontologies*

General Terms

Design, Human Factors

Keywords

Navigation, Knowledge Exploration, Semantic Metadata, Semantic Browsing, Exploratory Search, Annotation

ACMSE '10 April 15-17, 2010, Oxford, MS, USA.

1. INTRODUCTION

The prevailing paradigm for information retrieval and exploration on the Web is based on keyword search and document browsing. Under such circumstances, the user is likely to undertake the following operations:

- First, assemble a set of keywords deemed ideal for retrieving "good hits."
- Select documents based on title links and document summaries for each hit appearing in the Search Engine Results Page (SERP).
- Manual document inspection for relevance verification based on overlap between document content and information need.
- Finally, optional result aggregation and organization, commonly through bookmarking, saving, publishing etc.

This interaction sequence suffers various limitations. First, since *query reformulation* is the only recourse if no relevant results are found, multiple queries may need to be constructed before satisfactory results can be obtained. Second, the ability to navigate to surrounding and related contexts becomes restricted to pre-established anchors provided by page creators. For example, if an exploratory-minded user begins with the search phrase "Father of the Web," he may be unable to examine a related context such as the "World Wide Web Consortium," (the organization chaired by the said father), unless hyperlinks exist apriori from documents in the SERP to documents in the corpus containing the term "World Wide Web Consortium." This dependency between information reachability and hyperlinks could further become problematic in text collections devoid of hyperlinks altogether, such as Medline[1].

Another limitation that renders this paradigm somewhat impractical occurs when a user's information need is not well defined to begin with. For example, a user interested in

[1]Medline comprises more than 19 million citations for biomedical articles with links to full-text articles

the "Haiti crisis" may be moderately interested in many aspects of this broad event, ranging from "relief and recovery, economic impact, casualties, political climate, crime etc." Under such a scenario, the number of documents that must be examined before the user begins to narrow their window of interest using only hyperlinked browsing and query reformulation, could be substantial. As noted by Guha et al. in [7], *Research Search* can be aided by context-driven navigation that leverage Semantic Web techniques to facilitate the information retrieval task. Sheth and Ramakrishnan identified many limitations of the Fetch and Browse strategy, in the context of Research Search [12], coining the phrase "search-and-sift."

The main motivation for this work arises from the realization that users are ultimately interested in information, not in documents. We recognize that the information sought by users is commonly *embedded within* documents. The progression of Semantic Web standards and technologies now makes it possible to annotate documents and connect relevant information within and across documents. Such semantic annotations and interconnections open new possibilities for enhancing information retrieval and exploration.

In this work, we address some of the shortcomings of the search-and-sift paradigm, making the following specific contributions:

1. We present a novel information exploration paradigm, based on a data-centric model of information.

2. We show how to leverage background knowledge (i.e. semantic metadata) in the form of named entities and named relationships to create a viable alternative to hyperlinked document browsing.

3. We present a prototype implementation that is suitable for knowledge discovery through this data-centric paradigm for exploratory search.

4. We provide utilities that enable organization and aggregation of search results for publishing and data sharing.

In the next section, we present the overall the system design and architecture.

2. ARCHITECTURE

Figure 1 shows a diagrammatic representation of the system architecture. A significant component of the architecture is the background knowledge (i.e. domain model), since it provides entry points to information in the corpus and forms the supporting structure for exploring related contexts. In the context of this work, the background knowledge is a collection of triples composing a knowledge or instance base (KB). A triple is a ternary relation containing an entity pair and a relationship that expresses the link between them i.e. subject-predicate-object. Another important component of the architecture is the Spotter Module, responsible for connecting the Document Corpus to the KB by recognizing and annotating entity mentions in text. It is through annotations of entity mentions and associated background knowledge that the Semantic Browser is able to guide user navigation through the corpus. The following sections explain each component in more details.

2.1 Knowledge Base

In practice, background knowledge may be acquired in several ways. It may pre-exist on structured data sources such as MeSH [2], UMLS [4], DBpedia[1] etc or may be garnered from user-generated, simple data entry interfaces such as the Semantic MediaWiki [3]. Alternatively, facts may also be extracted directly from text, using supervised learning techniques [5] [9] as well as unsupervised learning techniques, such as those described by Ramakrishnan et al., [10] [11].

Our system has been developed in a highly modularized fashion and is therefore agnostic to the knowledge acquisition methods used to populate the KB. The only requirement is that the KB be expressed as triples (entity-attribute-value); preferably accessible remotely through Web Services following the RESTful design [6]. The current prototype system is connected to three (3) knowledge bases, the:

- Unified Medical Language System (UMLS) KB containing 5,232 entities and 16,540 triples, made available through the National Library of Medicine (NLM).

- Human Performance and Cognition Ontology (HPCO)[2] containing 15,742 entities and 22,298 triples

- Linked Open Data (LOD)[3], accessed remotely as a REST Web Service (`http://lod.openlinksw.com/sparql`).

2.2 Controlled Vocabulary

The Controlled Vocabulary is a collection of known entities names. It can be used to guide the identification of named entities that appear in text. Such a vocabulary can be constructed from entities in the knowledge base, for instance. For our purposes, three controlled vocabularies have been loaded in the current system; one containing almost 1 million entities from DBpedia, one containing all entities from the HPCO Ontology, and one containing 5,232 entities from UMLS. We discuss the selection and use of each in section 2.3.

2.3 Spotter Module

The Spotter Module is the component of the system for identifying and annotating named entities mentioned in text. The current spotter performs "exact label matching" between a sequence of tokens and entities present in the controlled vocabulary. This primitive form of entity identification has two major drawbacks. The first and obvious limitation lies in the overwhelming necessity for disambiguating polysemous named entities. For example, in the sentence *"the canon eos 40d is a 10.1-megapixel semi-professional digital single-lens reflex camera,"* it is discernible to the human reader that "canon" refers to a company brand and not to a city in Georgia, United States. However this distinction is obscure to the spotter. The second limitation of exact matching is the challenge of spotting complex entities, such as "canon eos 40d" in the aforementioned sentence. While Canon represents the electronics company, the string "canon eos 40d" represents a product (digital camera). Although

[2]HPCO is a funded project involving the Air Force Research Lab (AFRL) at the Wright-Patterson Airforce Base and the Kno.e.sis Center (`http://knoesis.org`)

[3]The LOD is a semantic web initiative to provide a repository of semantically connected datasets

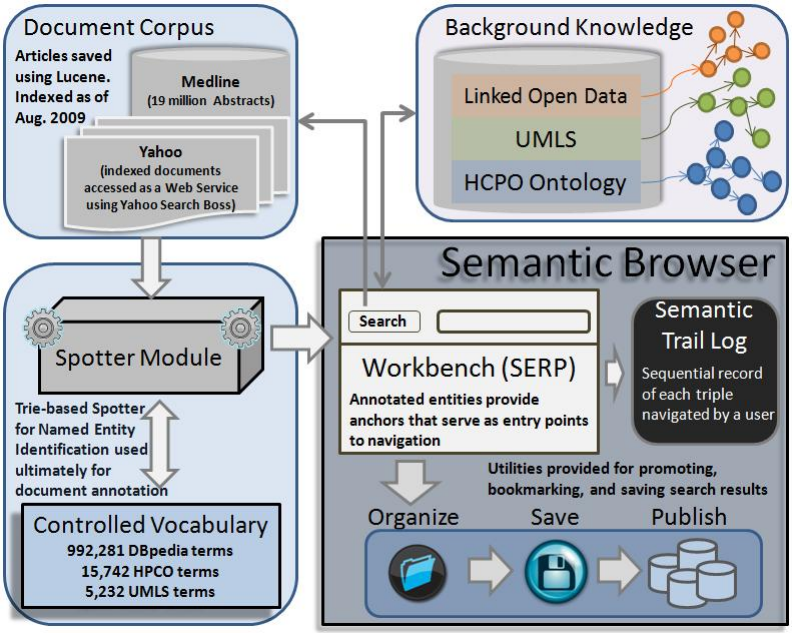

Figure 1: System Components and Architecture

the current implementation does not support disambiguation, we intend to adapt techniques from relevant work done by a number of researchers, including previous work at the Kno.e.sis center.

Our spotter utilizes a data structure called a "trie" (prefix tree), targeted for optimized search on longest prefix matches of labels in text. We build a trie with the entries in our Controlled Vocabulary and store it in main memory. At run time, we perform a search on the trie for the entries that match each token in our document summary, left-to-right. At each step, we select the entry with the maximum token overlap and restart the process with the next token after the overlap. While this simplistic technique is sufficient for this initial phase, more advanced techniques for spotting are warranted.

2.4 Document Corpus

The third component of the system is a collection of textual documents rich with information sought by the user. Spotting entities within such a corpus is a key task in providing entry points to navigation. The current prototype uses two corpora:

- the entire Medline corpus, containing 19 million abstracts as of August 2009, accessed as a RESTful Web Service, for which the UMLS and HPCO data dictionaries are used to spot entities, and the UMLS and HPCO KBs are used for navigation.

- search results from Yahoo! retrieved dynamically using the Yahoo! Search Boss Web Service API [4], for which the DBpedia dictionary is used to spot entities, and the Openlink LOD knowledge base[5] is used for navigation.

[4] `http://developer.yahoo.com/search/boss/boss_guide/`
[5] `http://lod.openlink.com`

2.5 Semantic Browser

The final component of the system is the Web pplication for contextual information navigation, organization and aggregation. Together with the *Semantic Trail Logs* and utilities for reuse and aggregation, the *Workbench* (SERP) creates the interface for the prototype tool, rendered as what is referred to as a "Semantic Browser."

The Semantic Browser is not a standalone web browser, such as Mozilla Firefox or Internet Explorer, but instead it is a Javascript application that runs within a standard web browser. The main components of the browser are:

- an input search box, which accepts keyword queries for keyword search.

- a workbench area for displaying annotated search results and for supporting context navigation.

- a semantic trails area, which maintains a record of triples traversed for aiding knowledge discovery.

- tools for organizing, saving and publishing search results.

We provide a complete walk-through of the system in section 3, showing a user interaction sequence with the various components, given a query. This example concretizes the navigation paradigm we propose in this work.

3. APPROACH

Figure 2 shows the interface in response to the query string "magnesium," using the Medline corpus and the UMLS KB for navigation. The system retrieves top ranked document summaries from the *Document Corpus* (i.e. Medline abstracts), which is hosted in a Lucene Index on a Kno.e.sis Web Server, and accessed as a REST Web Service. Each abstract is then spotted, using the *Spotter Module* to identify and annotate navigable entities present in the UMLS

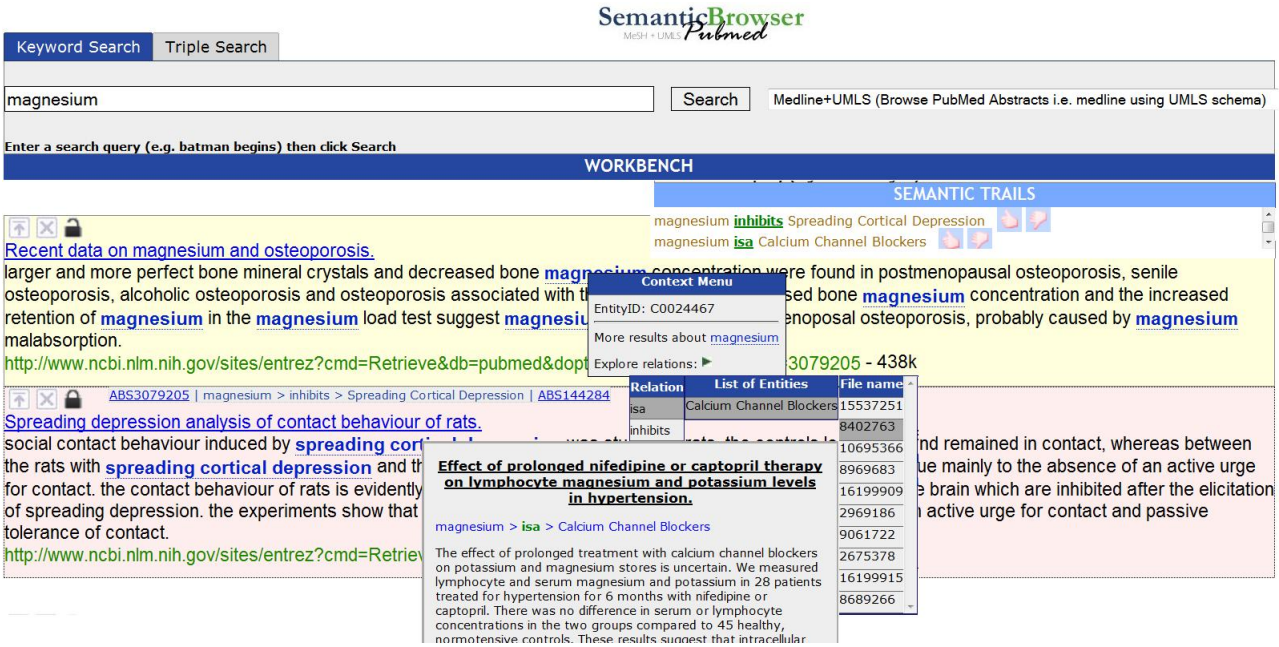

Figure 2: Snapshot of the User Interface

Controlled Vocabulary. In this example, "magnesium" and "magnesium deficiency" have been spotted and highlighted in blue boldface font.

Upon clicking an annotated entity, (e.g. magnesium) the system queries the *Background Knowledge* for a description of the entity (if available). Along with the entity description, the system retrieves and presents a list of relationships for which the entity (i.e. magnesium) is the subject. This relationship list is available through the "Explore relations" option. On mouseover Explore relations, the system populates the workbench with this list of relationships (i.e. isa and inhibits). On relationship mouseover (e.g. isa), the system presents a list of all entities in the object of any triple containing magnesium as the subject and 'isa' as the predicate. Hence, this highlighted example shows the triple *"magnesium* $\longrightarrow$ *isa* $\longrightarrow$ *Calcium Channel Blockers."* On object mouseover (i.e Calcium Channel Blockers) the system shows the document ids of abstracts containing the selected object (i.e Calcium Channel Blockers). Further, on document id mouseover, a snippet of the document is displayed for consideration by the user. Any document, if selected, becomes imported into the workbench immediately below the parent document containing the entity that began the navigation. Simultaneously, a record of the semantic trail traversed from subject-predicate-object is attached to the imported document as well as in the *Semantic Trail Log* (click the "Show Trails" menu option in the upper left corner of the browser). Finally, the user can reorganize the workbench by promoting, deleting and bookmarking selected documents based on their information need and the information contained in the documents. Bookmarked abstracts (shown in pink) remain in the workbench throughout the search session, remaining persistent in the SERP even after another query is executed. The final state of the workbench may be saved for publishing as a means of information sharing and reuse.

4. EXPERIMENTAL RESULTS

To demonstrate the effectiveness of this innovation we conducted a *quantitative study using subjective metrics* to assess user reaction to the system compared with PubMed (i.e. user interface for browsing Medline articles) and the Yahoo search engine for the same information need. We selected PubMed since it is devoid of hyperlinks and therefore likely to involve numerous query reformulations to retrieve the desired information. We selected Yahoo to show that our approach aptly addresses the search-and-sift limitations of a traditional search engine for Research Search. The pilot study was conducted with 13 users, all graduate students in Computer Science except two post-docs and one visiting researcher. We based our user study on Swanson's motivating scenario [13], summarized below. Each subject was provided the following instructions:

> "Dr. Swanson found a link between magnesium and migraine by manually examining the titles of abstracts of scientific literature. Use the semantic browser (`http://knoesis.wright.edu/trellis`), PubMed (`http://www.ncbi.nlm.nih.gov/sites/entrez`) and Yahoo search (`http://yahoo.com/`) to find this link. Fill in our evaluation form upon completion."

Figure 3 summarizes the results. Users were asked to rank systems on a [1 - 5] scale (1 - poor, 5 - excellent). The results are shown for each measure as a relative aggregated score. The sum of all scores for each user interface category is normalized by the user interface with the highest score. (e.g. "Interface Design" - Semantic Browser score: 40, PubMed score: 38, Yahoo score: 43, yields the fractions in the first row by dividing each score by the maximum 43).

The results above capture many important revelations about user perceptions of the system. It shows that users found the features of our system much more useful than

Evaluation Metrics	Search User Interfaces		
	Semantic Browser (Medline + UMLS)	PubMed	Yahoo
Interface Design	0.93	0.88	1.00
Useful Features	1.00	0.67	0.65
Motivation to Explore	1.00	0.58	0.65
Information Novelty	1.00	0.76	0.79
Effectiveness of Task outcome	1.00	0.65	0.80
Required Cognitive Load	1.00	0.60	0.64
Overall Satisfaction	1.00	0.62	0.78

Figure 3: Evaluation Results from Usability Study

the other two systems and were very motivated to continue exploration. It also shows that the amount of cognitive load required to use the browser was lower, since the system presents context to the user in the form of relationships as they browse. Users also found the amount of novel or additional information rendered based on their initial information need was greater in the browser. In summary, users were most satisfied with our system. Granted that our evaluation sample is limited, the overall conclusion is that our proposed paradigm is promising and warrants further research.

We do acknowledge that the ability to conduct user studies of this nature is limited by the unavailability of good evaluation metrics for browsing. The traditional precision and recall metrics used in information retrieval may not be applicable here, because we intend to measure how quickly users satisfy their information need, not necessarily through documents. The navigated triples themselves may be sufficient to provide desired information to satisfy an information need. Marchionini [8] notes that in recent years researchers tend to focus on the design and features of new systems and interfaces to support exploratory search activities, not on their evaluation. And while subjective measures such as user satisfaction, engagement, information novelty and task outcomes are also important, interaction behaviours, cognitive load, and learning have been suggested as better indicators of effectiveness of Exploratory Search Systems. We included many of these metrics as our evaluation parameters in support of the proposition.

5. FUTURE WORK

In spite of positive initial results, many important tasks remain to be enhanced to ensure greater accuracy and system adaptation of a wider data and corpora spectrum. For example, more reliable techniques for spotting, beyond exact label matching are warranted. Context-aware entity spotting that takes into account words in the surrounding neighbourhood of potential entities appears to be a viable alternative. Additionally, a mechanism for recognizing complex entities in text is also necessary in order to accurately anchor entry points into navigation. Further still, it may be possible to enhance navigation by implementing some form of relationship ranking, based on perceived relationship relevance to user interests, to better align contexts to user's train of thoughts. Finally, we consider more robust evaluation by comparison with standard information retrieval metrics as a feasibility study.

6. CONCLUSION

We prototyped a novel information exploration system that superimposes a trellis[6] of entities and relationships over a corpus of scientific literature (i.e. Medline). Our browser better supports navigation of related contexts through semantic metadata in the form of annotated entity mentions in text and their relationships obtained from various knowledge bases. We also provide utilities for bookmarking and document reorganization, maintaining a semantic trail of user "train of thoughts." Analysis of such trails potentially leads to newly discovered knowledge over text where hypotheses can be later corroborated scientifically. We therefore believe our system to be a novel data-centric paradigm for information exploration that goes beyond the document-centric model of the traditional search-and-sift paradigm.

7. ACKNOWLEDGEMENTS

We thank Cartic Ramakrishnan for the conceptualization of the first Semantic Browser[7], and for his suggestions during the development of the current system. Special thanks to many students who contributed to the development of the Semantic Browser, including Bilal Gonen, Aditya Dhoke, Wesley Workman, Rodrigo Gama and Guilherme de Napoli. This research is supported by research grants from the Human Effectiveness directorate of the Air Force Research Lab, WPAFB and the National Science Foundation under Award No. 071441 to Wright State University and No. IIS-0325464 to University of Georgia titled "SemDis: Discovering Complex Relationships in the Semantic Web." Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

8. REFERENCES

[1] Dbpedia. `http://dbpedia.org/About`, January 2010.

[2] Mesh database. `http://www.ncbi.nlm.nih.gov/mesh`, January 2010.

[6] A trellis is an intricate architectural structure for supporting plants and vines

[7] http://lsdis.cs.uga.edu/projects/semdis/SemanticBrowser/

[3] Semantic media wiki. `http://semantic-mediawiki.org/wiki/Semantic_MediaWiki`, January 2010.

[4] Unified medical language system (umls). `http://www.nlm.nih.gov/research/umls/`, January 2010.

[5] Daniel M. Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. Nymble: High-performance learning name-finder. In *In Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 194–201, 1997.

[6] Roy T. Fielding. *Architectural Styles and the Design of Network-based Software Architectures*. PhD thesis, University of California, Irvine, 2000.

[7] R. Guha, Rob Mccool, and Eric Miller. Semantic search. In *International World Wide Web Conference*, pages 700–709. ACM, 2003.

[8] Gary Marchionini. Editorial: Reviewer merits and review control in an age of electronic manuscript management systems. *ACM Trans. Inf. Syst.*, 26(4), 2008.

[9] Andrew Mccallum and Wei Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. 2003.

[10] Cartic Ramakrishnan, Pablo Mendes, Shaojun Wang, and Amit Sheth. Unsupervised discovery of compound entities for relationship extraction. pages 146–155. 2008.

[11] Cartic Ramakrishnan, Pablo N. Mendes, Rodrigo A. Gama, Guilherme C. Ferreira, and Amit P. Sheth. Joint extraction of compound entities and relationships from biomedical literature. In *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT '08. IEEE/WIC/ACM International Conference on*, volume 1, pages 398–401, 2008.

[12] Amit P. Sheth and Cartic Ramakrishnan. Relationship web: Blazing semantic trails between web resources. *IEEE Internet Computing*, 11(4):77–81, 2007.

[13] D. R. Swanson. Migraine and magnesium: eleven neglected connections. *Perspect Biol Med*, 31(4):526–557, 1988.