Idea Navigation: Structured Browsing for Unstructured Text

Robin Stewart
MIT CSAIL
32 Vassar St.
Cambridge, MA 02139 USA
stewart@csail.mit.edu

Gregory Scott
Tufts University
161 College Ave.
Medford, MA 02155
greg.scott@tufts.edu

Vladimir Zelevinsky
Endeca
101 Main St.
Cambridge, MA 02139 USA
vzelevinsky@endeca.com

ABSTRACT
Traditional interfaces for information access do not fully support queries that rely on semantic relationships between terms. To better support such queries, we introduce a system that automatically extracts subject-verb-object concepts from unstructured text documents and dynamically presents them to the user as navigable refinements. This approach, which we call "idea navigation," makes subject-verb-object querying as simple as selecting successive refinements. It also supports exploratory search by providing a view of the most common ideas in the current result set. First-time users of a prototype system successfully used idea navigation to solve realistic search tasks, demonstrating its effectiveness.

ACM Classification Keywords
H5.2 **[Information interfaces and presentation]** User Interfaces—*Interaction styles*; I.2.7 **[Artificial Intelligence]** Natural language processing—*Text analysis*.

Author Keywords
Exploratory search, faceted browsing, information retrieval.

INTRODUCTION

It is well established that the omnipresent search box is insufficient for supporting many common information-seeking tasks [1]. To improve the situation, many commercial websites now also provide "faceted browsing" interfaces, which give users the ability to view only those results matching a particular set of metadata [2]. For example, one might search for "televisions" and then narrow the results by clicking on facet refinements such as "flat screen" or "$1500-$3000". This type of interface is particularly useful for exploratory search tasks, where users may not know how to define a priori the best query to solve their task – whether because they don't know in advance what information is available in a particular collection or because they cannot anticipate which keywords would best describe their desired results.

Although a number of interfaces have been designed to further expand support for various types of exploratory search [3], the refinement options provided by these interfaces are limited to human-assigned metadata and keywords extracted from text. This limitation is acceptable for searches that happen to be a conjunction of the available keywords or metadata, as in a search for flat-screen televisions that cost $1500-$3000. But what if we are looking for something more abstract or subjective? For example, we may want to find "historical events that are interesting in the present context" or "quotations that one might find controversial." The available metadata (such as "author" or "subject") is unlikely to be useful for these tasks.

Further, even when relevant facets do exist, the query may depend on a relationship *between* facets. Consider a search for campaign proposals made by Hillary Clinton. It would

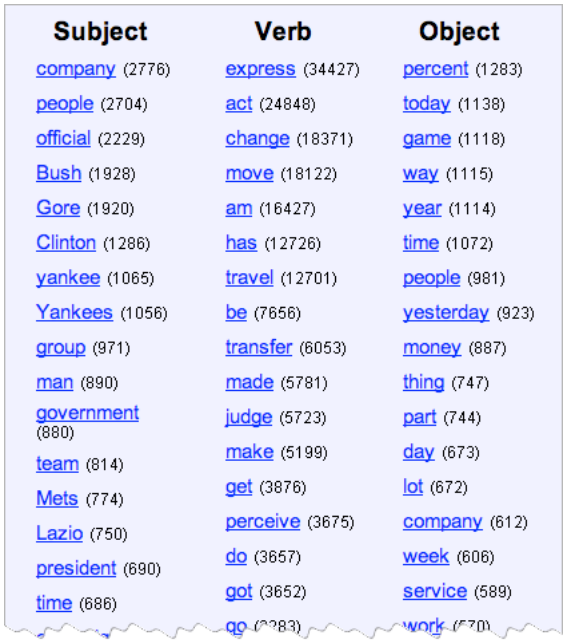

Figure 1. The idea navigation interface summarizing ~9000 news articles from October 2000.

CHI 2008, April 5–10, 2008, Florence, Italy.

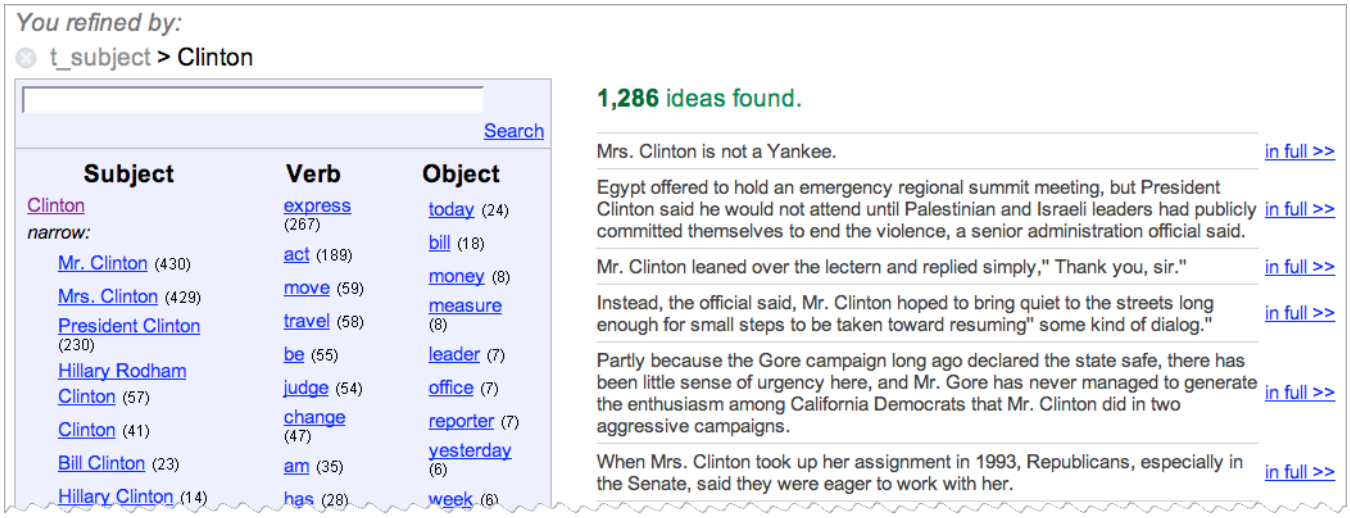

Figure 2. The full system interface after clicking on "Clinton". Further refinement options are on the left. Sentences from the document collection that have a Clinton as their subject appear on the right.

not be sufficient to look for documents that contain the terms "Hillary Clinton" and "proposals", because the result set will contain many documents in which Clinton is not the person doing the proposing. This type of query is particularly common in scientific, legal, and patent searches (e.g.: find molecules that target a particular cell; find inventions that burn solids for locomotion). Current interfaces make such search tasks awkward at best, since users need to scan through the list of matching articles for passages that might indicate relevance.

To better support these types of queries, we have developed an interface for searching natural language documents, which takes advantage of the linguistic information provided by English sentence structure. Our system initially scans every sentence in the document collection to extract sets of words (subject–verb–object) that indicate the presence of an idea, such as "Clinton–proposed–reforms." It groups similar terms together under broader categories so that they can be summarized more effectively. The system then dynamically aggregates all of the ideas and presents them to the user as navigable refinements (Figure 1). This approach, which we call "idea navigation," gives users the ability to perform complex subject-verb-object concept filtering in a manner that is as easy to understand as standard faceted browsing interfaces: navigation is performed by simply selecting successive refinements. As a bonus, the list of refinements provides a helpful view into the most common ideas in the current result set.

IDEA NAVIGATION

We demonstrate idea navigation by answering one of the questions above: What did Hillary Clinton propose in previous campaigns? Our prototype system contains all ~9000 articles published by a popular news source in October 2000, when, Clinton was running for the Senate. We first select "Clinton" in the Subject column, revealing a variety of narrow terms such as "Mr. Clinton" and "President Clinton" (Figure 2). We choose "Mrs. Clinton" to refine the results to sentences with Mrs. Clinton as their subject. Selecting the broad action "express" in the Verb column refines the result set to things that Mrs. Clinton said (Figure 3). This results in 117 matching ideas (from 46 different documents), so we select the narrow verb "propose", which narrows down the result set to five sentences. All five provide answers to our query. One is: "Mrs. Clinton, for example, has proposed federally financed scholarships for college students who commit to teaching."

We chose to represent ideas as subject–verb–object triples

Figure 3. Idea navigation facets after choosing "Mrs. Clinton" as the Subject and "express" as the Verb.

because this representation has been used successfully in question answering systems (to help answer focused questions such as "Whom did Hillary Clinton marry?") [4] and in advanced search box functionality [5]. This representation is also the basic underpinning of the semantic web's Resource Description Framework. Numerous interfaces for searching such semi-structured data have been built, including ESTER [6], which proactively displays refinement options that it considers relevant. However, to our knowledge, our system is the first to present the ternary representation as a faceted browsing interface.

In addition to its browsing capabilities, our prototype system also includes a search box, which lets users perform a search over all ideas in the collection. Search and idea navigation refinements can be used together to narrow the results. We expect that commercial information retrieval interfaces will include fully-featured search and faceted browsing of metadata; meanwhile, the intention of this prototype is to provide enough supporting functionality to allow users to answer real questions so that we can evaluate the usability of the idea navigation module.

The precision of our extraction process is very high: virtually every extracted triple correctly maps to a true subject-verb-object structure in the source text. The recall is lower because not every subject-verb-object structure is found; since this does not affect the interface's accuracy, we do not expect it to influence our evaluation results.

EVALUATION

We carried out a formative evaluation to test whether users would (a) understand the idea navigation interface after a brief introduction, (b) choose to use that interface when given the option alongside a standard search box, and (c) successfully complete tasks with its help. We recruited 11 participants, who ranged in age from 19 to 30 and performed Internet searches from 5 to 150 times per day. Most of the subjects were university students and staff; none were professional researchers, nor had any of them used idea navigation previously.

We began each session with a demonstration similar to that given above: we first attempted the "Hillary Clinton proposed" search using the standard search box on a popular news site, and then used idea navigation to locate the relevant information. We also answered any questions about the interface's behavior and gave participants up to 3 minutes to explore it.

We then asked each user to perform a set of tasks inspired by the kinds of queries we suspect search boxes and faceted metadata do not fully support: queries that concern specific relationships or subjective judgments. All tasks could be accomplished using idea navigation refinements only, but the search box was also available (Figure 2).

Task 1. In October 2000, a Yankee pitcher named Roger Clemens threw a bat at an opposing player. Find the opposing player's name.

Plenty of articles contain Clemens, bats, throwing, and opposing players, but this question requires specific relationships between these elements. All participants completed the task in less than three minutes, and most completed it in less than one minute; but no two participants used the same path to find the answer. Most instinctively started with a search, such as "Roger Clemens", "bat", or "Yankee"; but then progressed to using refinements (including Subject: "Clemens", Verb: "throw", and Object: "bat") until a sentence with the answer appeared high in the results.

Task 2. Find something about George W. Bush in Oct. 2000 that is interesting to you in the context of the modern day.

This task was intended to be open-ended; however, refining by Subject: "Bush" would clearly be a helpful approach. Only half the participants made this refinement initially, but all except one did so eventually during the task.

Task 3. Find a quotation that could be considered controversial or offensive.

A prototypical user session for this task was as follows:

User searches for "controversy"... scans sentences... starts over.
Refines by Verb: "express" > "say"... scans... starts over.
Searches for "offensive"... scans... starts over.
Says: "I'm thinking there's nothing in the corpus that it says is offensive..."
Experimenter: "Find something *you* interpret as controversial or offensive."
Searches for "race black"… no results.
Searches for "african american"
Refines by Verb: "resegregate"... and reads the article.

Many users initially tried searching for "controversy", "quote", "comment", etc. – even when the experimenter used different synonyms to describe the task. However, most of the time this failed to produce useful results because very few controversial quotes are actually labeled as such in the article. Therefore, users turned to the idea navigation refinements, which let them find subjects and verbs that they thought would lead to controversy. Some of the refinements found were "criticize", "denounce", "attack", and "condemnation".

DISCUSSION

Presumably due to the dominance of keyword search interfaces today, most subjects had a clear initial bias towards formulating the tasks as keywords in the search box, such as "roger clemens throw bat" or "offensive statement." Even so, subjects consistently and successfully used the idea navigation refinements to improve upon their search box results. Most participants increased their usage of idea navigation over time, in total performing 100 idea navigation refinements versus only 61 searches. All tasks were successfully completed and 79% were completed with idea navigation as the final search step.

These results demonstrate that the addition of idea navigation indeed provides a significant improvement over stan-

dard search boxes for the types of queries we tested. To convince ourselves that this finding was not due simply to the limitations of our search box implementation (which searched triples rather than full text and did not use ranked retrieval), we asked some of the participants to also perform Task 3 with the real search box at a news site. These users had a similar level of difficulty finding controversial statements as they had with our search box implementation.

More studies are necessary to compare idea navigation against alternative interfaces such as metadata facets and tag clouds. Although all of these interfaces are useful for general exploratory search, only idea navigation supports queries that require specific relationships between terms. We believe that this ability gives idea navigation a substantial advantage for many applications.

IMPLEMENTATION

Our system uses a pre-processing phase to extract the subject–verb–object triples from the document set. We lack the space here to examine all of the natural language processing considerations; see [4] for a fuller description of a similar extraction process. The steps we follow are:

1. **Parsing.** We use the Stanford Parser [7] to generate parse trees and label each word with a part of speech.
2. **Pronoun resolution.** We modified a LingPipe [8] module to perform anaphora resolution, translating pronouns such as "she" or "it" into the entities they refer to. For example, "She has proposed taxes" would become "Mrs. Clinton has proposed taxes" if "Mrs. Clinton" was the last mentioned female living entity.
3. **Triple extraction.** Using the parse trees from step 1, the system recursively looks for patterns that identify connected subjects, actions, and objects. It handles sentences structured like the above example, as well as passive voice statements such as "taxes proposed by Mrs. Clinton"; both become the triple "Mrs. Clinton–propose–taxes." Ideas with only two components are also accepted: "Hillary–competed."
4. **Cleanup.** This step discards triples that carry little information, such as: "Hillary–is".
5. **Aggregation.** The system groups similar triple-components together. For subjects and objects, it groups by the head noun of the noun phrase: "Mrs. Clinton" is under "Clinton". For verbs, it uses the topmost hypernym in the WordNet [9] hierarchy: "propose" is under "express".

We store the extracted ideas in an Endeca database [10] since it is optimized for generating facet refinements on the fly. A web application queries the database to retrieve the most frequent subjects, verbs, and objects (along with the sentences they came from) for the current result set. It then displays these to the user as refinement options along with a list of sentences from which they were extracted.

CONCLUSIONS AND FUTURE WORK

In this paper, we demonstrated how document search interfaces could be enhanced by using the faceted browsing interaction style with the subject–verb–object representation of ideas. Our user study demonstrated that such an interface is understandable to first-time users and useful for solving search tasks that are poorly supported by existing systems.

More elaborate systems can be built that extract more information from natural language text. Possibilities include: increasing the number of sentence structures that the system understands; trying different ways of grouping the idea components; and exploring alternate idea representation schemes that better handle adjectives and prepositional phrases. We also plan to integrate idea navigation into a full-featured search interface with a large health science document set, allowing us to perform a more extensive, comparative user study that examines user performance when various search components are available.

ACKNOWLEDGMENTS

Many thanks to Daniel Tunkelang, Blade Kotelly, Shiry Ginosar, Boris Katz, Rob Miller, Max Van Kleek, Michael Bernstein, Adam Marcus, Joyce Wang, Karen McManus, and David Karger for their invaluable advice and feedback.

REFERENCES

1. White, R. W., Kules, B., Drucker, S. M., schraefel, m. c. (2006) Supporting Exploratory Search, Introduction, Special Issue. In *Communications of the ACM* 49(4), pp. 36-39.
2. Hearst, M., English, J., Sinha, R., Swearingen, K., and Yee, K.-P. (2002) Finding the Flow in Web Site Search. In *Communications of the ACM* 45(9), pp. 42-49.
3. Wilson, M., schraefel, m.c., White, R. (2007) Evaluating advanced interfaces using established information-seeking models. Technical Report, School of Electronics and Computer Science, University of Southampton. http://eprints.ecs.soton.ac.uk/13737/
4. Katz, B., Lin, J. (2003) Selectively Using Relations to Improve Precision in Question Answering. In *Proc. EACL 2003 Workshop on Natural Language Processing for Question Answering*.
5. http://www.powerset.com/
6. Bast, H., Chitea, A., Suchanek, F., Weber, I. (2007) ESTER: Efficient Search on Text, Entities, and Relations. In *Proc. SIGIR 2007*, ACM Press, pp. 671-678.
7. http://nlp.stanford.edu/software/lex-parser.shtml
8. http://www.alias-i.com/lingpipe/
9. http://wordnet.princeton.edu/
10. http://www.endeca.com/