Semantic Annotation Based Exploratory Search for Information Analysts

Jae-wook Ahn, Peter Brusilovsky, Jonathan Grady, Daqing He
School of Information Sciences, University of Pittsburgh

Radu Florian
IBM TJ Watson Research Center

Abstract: The system presented in this article aims to improve information access through the use of semantic annotation utilizing a non-traditional approach. Instead of applying semantic annotations to enhance the internal information access mechanisms, we use them to empower the user of an information access system through an innovative named entity-based user interface – NameSieve. NameSieve was built to support an intelligence analyst during the process of exploratory search, an advanced type of search requiring multiple iterations of retrieval interleaved with browsing and analyzing the retrieved information. The proposed approach was implemented in the NameSieve system so that the system can transparently present a summary of search results in the form of entity "clouds." Therefore, these clouds allow the analyst to further explore the results in a novel manner, acting together as a faceted browsing interface. We ran a user study (with ten subjects) to examine the effect of NameSieve, and the study results reported in the paper demonstrate that this new way of applying semantic annotation information was actively used and was evaluated positively by the subjects. It enabled the subjects to work more productively and bring back most relevant documents.

Keywords
Semantic annotation, exploratory search, named entity, user interface, empirical study.

1. Introduction

A range of modern semantic annotation approaches makes it possible to annotate documents with higher-level semantic features from ontological concepts to named entities (names of people, places, organizations, etc.). Many researchers argue that semantic features are able to better model essential document content, and that their application can improve the user's ability to find and access the right information at the right time. A number of projects confirmed the potential of semantic annotations, applying them at different stages of the information processing and retrieval mechanisms (Demner-Fushman & Oard, 2003; Mihalcea & Moldovan, 2001; Wu, He, Ji, & Grishman, 2008). The work presented in this paper follows the research stream on improving information access through the use of semantic annotation, yet it attempts to reach the same goal from an alternative direction: empowering the user of an information access system through an innovative named entity-based user interface for exploratory search.

Exploratory search is described by Marchionini as a type of search "beyond lookup", such as *search to learn* and *search to investigate*. Exploratory search assumes that the user has some broader information need that cannot be simply solved by a single "relevant" Web page, but requires multiple iterations of search/analysis interleaved with browsing and analyzing the retrieved information. The research on supporting exploratory search attracts more and more attention every year for two reasons. On one hand, the number of users engaged in exploratory search activities is growing (Marchionini, 2006). With the exponential growth of information available on the Web, almost any user performs searches "beyond

lookup" even to plan a vacation or choose the "best" digital camera. Moreover, some classes of users, such as intelligence analysts, perform multiple exploratory searches every day as a part of their job. On the other hand, traditional search systems and engines working in a more simple mode of "query → list of results" provide very poor support for exploratory search tasks (Marchionini, 2006). Users have great difficulty formulating effective queries when they are unsure of their information needs. The challenge is compounded when the user is trying to make sense of search results presented only as a linear list.

Our team investigated the issue of exploratory search in the context of the DARPA GALE (Global Autonomous Language Exploitation) project. Our goal was to develop a more effective information distillation interface for intelligence analysis. We initially focused on personalized search, expecting that adaptation to an analyst's global task (beyond a single query) would enable our system to produce and bring better results to the analyst's attention. However, user studies performed by our team to evaluate personalized search interfaces (Ahn, Brusilovsky, He, Grady, & Li, 2008) convinced us that traditional personalized search is not sufficient to provide the proper level of support in an information exploration context. First, an extensive analysis of search logs produced by intelligence analysts revealed that query formulation is a major problem. The analysts struggled to bring hidden relevant documents to the surface by repeating various combinations of just a few of the most obvious query terms, while more powerful and less evident terms were never discovered. Second, on several occasions the analysts asked for an interface that provides "more transparency" and "more control" over the search process. Unfortunately, traditional personalized search offers no support for query formulation and no user control over the process. Personalization starts with an already submitted query and works as a black box, which produces a user-adapted list of results without direct user involvement. Inside this black box, the personalization engine applies a user profile either to generate query expansion or to reorder search results (Micarelli, Gasparetti, Sciarrone, & Gauch, 2007).

The work presented in this paper attempted to address these problems by exploring an alternative approach to support users in their exploratory search tasks. Instead of using artificial intelligence (AI) for query expansion and results reordering, we attempted to build an information exploration interface that enhances the user's own abilities in all three tasks: query formulation, query expansion, and re-ranking of the results. The key idea of the proposed approach is the application of named entities (NEs), a popular kind of semantic annotation, to present the *aboutness* of the search results to the users and to allow them to manipulate and explore these results.

The proposed information exploration approach was implemented in NameSieve, an information exploration interface for intelligence analysts and evaluated in a controlled user study. The following sections of this paper presents a description of the NameSieve interface along with a detailed account of how it was built and the results of the user studies. We also review similar work and discuss the potential of integrating the new information exploration interface with our other personalized search approaches.

2. Named Entities in Information Retrieval

As a semantic category, named entities (NEs) act as pointers to real world entities such as locations, organizations, people, or events (Petkova & Croft, 2007). Because NEs can provide much richer semantic content than most vocabulary words, they have been studied extensively in various language processing and information access tasks. NEs have been viewed as alternative information for indexing. Mihalcea and Moldovan (2001) discussed the idea of using NEs for indexing document content, and they found that the size of the index could be greatly reduced while relevant documents still can be retrieved.

As the most common type of out-of-vocabulary terms that do not have translations in the dictionary, the translation of NEs have been treated as a serious problem in dictionary-based Cross-Language Information Retrieval (Oard, 2002). Demner-Fushman and Oard examined the effect of out-of-vocabulary terms, where the majority are NEs, in CLIR through artificial degradation of the dictionary coverage (Demner-Fushman & Oard, 2003). They find that the performance can decrease by as much as 60% when NEs are removed from the translations. Through review of the search topics and retrieval systems in several years of Cross-Language Evaluation Forum (CLEF) experiments, Mandl and Womser-Hacker evaluated the NEs in those topics and their effects on CLIR (Mandl & Womser-Hacker, 2005). They found that the majority of CLEF topics contain at least one NE, and NEs often make retrieval topics relatively easier to obtain than those topics that do not have any NEs. Of course, their assumption is that reasonable translations can be found for these NEs. Wu and others further examined the effect of special handling of NEs and their translations using IE technology in task-based multilingual information exploration, and found that significant impact on retrieval effectiveness can be achieved with high quality translations of NEs (Wu, et al., 2008).

As the research and practice in information retrieval moved from classic ad-hoc retrieval scenarios to new challenges and applications, the roles of NEs have been considered more often for specific tasks. For experiments on topic detection and tracking, NEs have been used extensively for modeling the essential features of seminal events and for differentiating between new events and existing ones (Kumaran & Allan, 2004), as well as for detecting novelty in documents and events (Yang, Zhang, Carbonell, & Jin, 2002). In terms of question answering and multilingual question answering, NEs also are the essential information for representing the needs behind the questions. Pablo-Sanchez, Martınez-Fernandez, and Martınez (2005) reported on multilingual NE processing in cross-lingual question answering and in web cross-language information retrieval. Pizzato, Molla, and Paris (2006) proposed using the extracted NE in pseudo relevance feedback for question answering. Although they did not obtain significant improvement by using NEs, they found that the causes are more related to the retrieval measures used in question answering. Khalid, Jijkoun, and Rijke (2008) talked about the effect of normalizing NEs in question answering, and found that even very simple normalization of NEs have a clear impact on the retrieval and answering tasks.

Compared to the related work in the literature, our work is based on the insight that NEs are semantically richer components for modeling than keywords. Therefore, our NameSieve system extensively uses NEs to represent the content of returned documents. However, our research focuses not on indexing or ranking algorithms, but on the support that NEs can provide in the users' sense-making process. In NameSieve, automatically extracted NEs, categorized into who (people), where (location), when (time) and what (events), are displayed along with the returned documents so that the essence of those documents can be quickly and flexibly explored by the users.

3. NameSieve: Named Entity-based Information Exploration System

The key idea behind NameSieve, our NE-based information exploration interface, is to extract NEs from the documents returned by the user's query and display them to the user (Figure 1). This idea offers several benefits. First, the search results become more transparent to the user: the most critical information (in the form of NEs) contained in hundreds of retrieved documents is brought to light. This helps users to *make sense* of the search results. Second, by showing the main NEs related to the user's original search terms, the system uncovers critical people, locations, and organizations relevant to the

users' tasks. Visualization allows users to immediately take the main NEs into account for query expansion and formulate new queries.

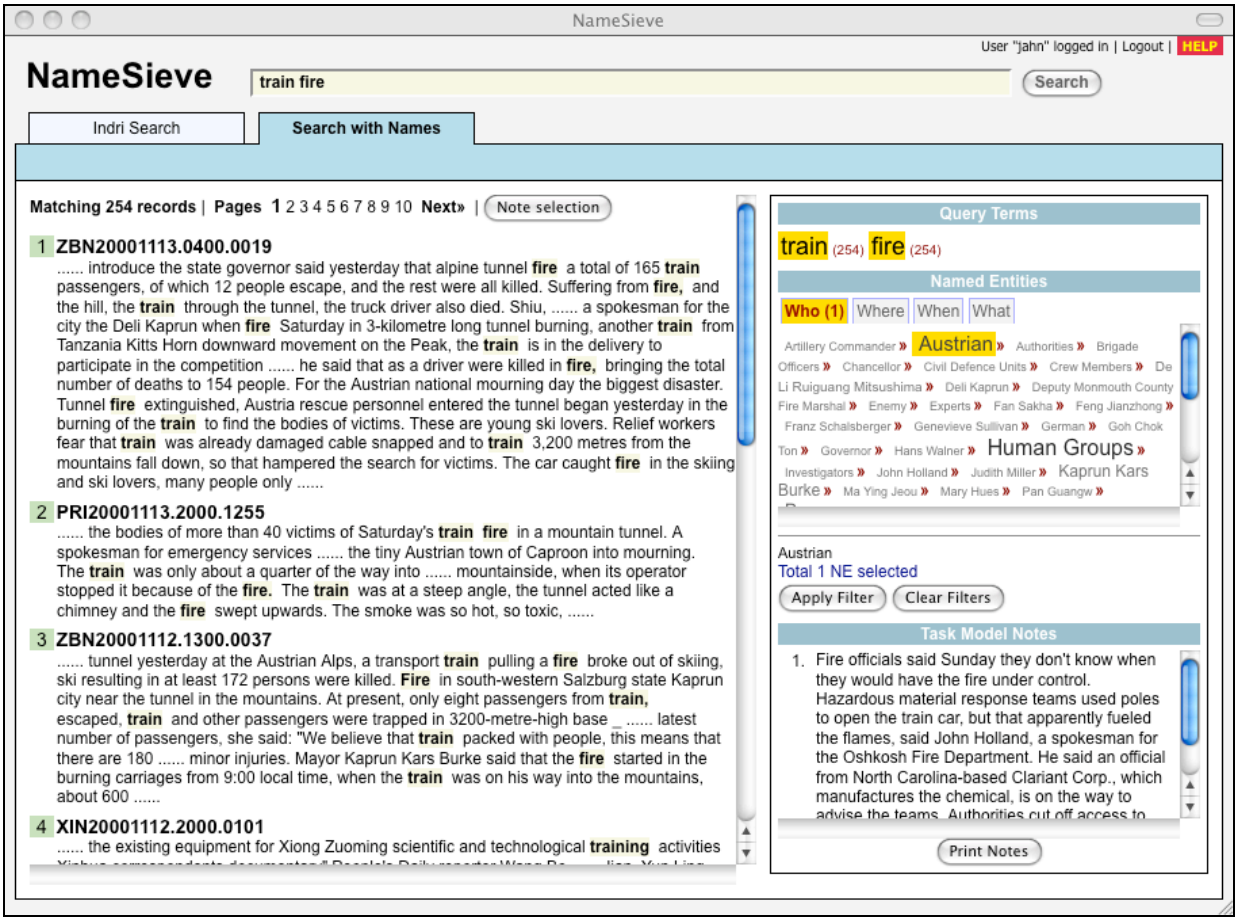

Figure 1: NameSieve Interface. Documents retrieved in response to *train fire* query are shown on the left. Named Entities extracted from these documents are shown on the right, at the top of the Control Panel. The user selected the NE *Austrian* and prepared to filter the results using *Apply Filter* button.

The second important idea is to complement the transparency achieved by NE extraction with user control. The list of extracted NEs in our system is not just a passive display, but an interface for instant query expansion and re-ranking of the retrieved results. The workflow supported by NameSieve is the following:

(1) User starts a new search by entering an initial query.

(2) The system retrieves documents using a traditional ad-hoc retrieval engine.

(3) The system processes the set of retrieved documents, extracts any NEs, and organizes them by their *prominence* in the list of results.

(4) The system displays the list of retrieved documents along with the organized list of extracted NEs.

(5) The user explores the presented documents and NEs. During this process, the user can *select* one or more interesting NEs as well as the original query terms.

(6) Selected NEs can be instantly added to the original query for a new search. In this case, the process begins again from step (1). Alternatively, the user can use selected NEs to *post-filter* existing search results, whereby the process moves to the next step.

(7) Given the selected NEs and search terms, the system updates the current list leaving only those of the originally retrieved documents that contain all selected items (query terms or NEs). The ranking of documents is now determined by their relevance to the selected items. Since this re-filtering reduces the number of retrieved documents, it also affects the set of associated NEs, which is now re-processed. The process restarts with step (4).

We used Indri for the baseline search engine in step (2) and implemented our own transparent Boolean filtering on step (7). We also used an advanced NE extractor mechanism developed at the IBM TJ Watson Research Center. Our experience demonstrated that both the quality of NE extraction and the organization of the interface are critical to making this idea work (see section 4 for more details).

Figure 1 shows our second-generation NameSieve's interface with an example taken from one of the study tasks (train fire at a ski resort). The user starts with a query "train fire". The system retrieves a large number of documents and immediately applies the default Boolean post-filtering, returning 254 documents containing both "train" and "fire". The matching documents are presented in a traditional style: 10 documents per page with document titles and surrogates generated using the sentences containing the user's query terms. Each term in the surrogates is highlighted, acting as a clue to help users understand why the corresponding document was retrieved by the baseline search system.

The user can operate with these results using the control area on the right hand side of the screen, which contains three panels: Query Term Panel, Named Entity Panel, and Notebook Panel. The Query Term Panel shows each term in the current query accompanied by the number of documents in the result list containing the respective term. Users can turn a filter on (highlighted in yellow, the default state) or off by clicking on a term. When a query term filter is turned on, the document list is updated to filter out all documents *not containing* the term. When a term filter is turned off, all relevant documents will be shown whether or not the term exists in a document. For example, if a user turns off the filter for the query term "fire", the new result list increases to 643 documents. The number of documents increases because the Boolean post-filtering was reduced from two terms ("train AND "fire") to one ("train"). The updated number of documents is displayed again, next to the term in the Query Term Panel.

The Named Entity Panel shown in Figure 2 is the core feature of the system. The system extracts and displays NEs from the list of documents on the left hand side of the interface. The NEs are organized into 4 tabs according to their types. The size and color of the displayed NEs are determined by their frequency. More frequently occurring NEs in the retrieved documents are rendered in a larger font and clearer color than less frequent ones. Unlike the query terms, whose filters are initially activated by default, NEs remain unselected waiting for the users to examine and select them based on the user's preference. When the NE filter selection is complete, the user clicks the "Apply Filter" button, and the system returns an updated document list. The updated list is post-filtered from the original list and includes only the documents that contain all of the selected names. This post-filtering process is done immediately on the entire list of documents retrieved from the previous session.

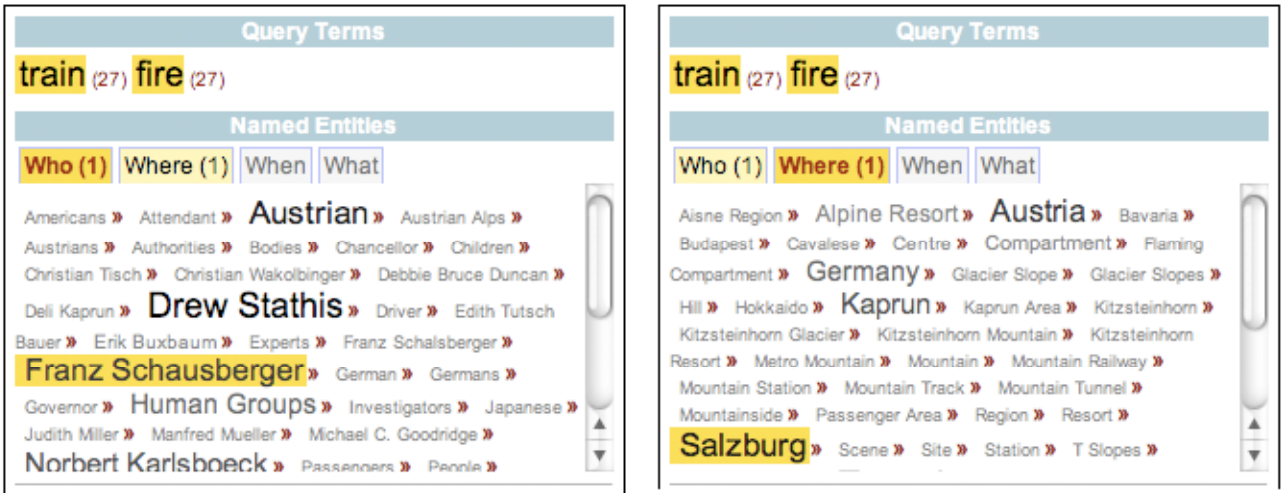

Figure 2: Named entity exploration interface
"Franz Schausberger" selected in the *Who* tab (left) and "Salzburg" selected in the *Where* tab (right)

Figure 2 shows an example of NE manipulation. Starting from the situation displayed in Figure 1, the user examines the NE list, selects the important location name "Salzburg" and clicks "Apply Filter" to narrow down the currently-retrieved list. When the filter is applied with "Salzburg", the number of documents in the list is reduced to 27, and the list of NEs is updated accordingly. The user examines the updated NE list and decides to learn about the connection between the Salzburg governor, Franz Schausberger, and the train fire. After selecting the NE "Franz Schausberger" (Figure 2) and applying the filters again, only 3 documents remain in the list to be examined in details by the user. The selected filters can be turned off again anytime, so that the search process using the NE filters is as flexible as possible.

To help users remember which NE filters are turned on within the four tabs, the number of selected NEs is displayed and the tab background changes to yellow. The label of the active tab, *Who*, in Figure 2 is rendered in red (foreground) and dark yellow (background), because the user selected the NE "Franz Schausberger". On the *Where* tab label, we can see that there is another selected name, a location name "Salzburg". In order to distinguish itself from the active tab, the background is rendered in light yellow. Below the box, all selected NEs are displayed in a smaller font size followed by the count, giving the user an overview of the exploration process outcomes.

4. Named Entity Extraction and Processing

As we found out during our work on the project, both the power of the mention detection stage and the quality of the post-processing stage are vital to the success of the overall approach presented in this paper. While the intent of our approach was clear from the very beginning, we had to explore several detection mechanisms, go through several major refinements of the post-processing pipeline, and run two user studies to achieve a quality of NEs which was meaningful for the users and which can significantly impact their work.

The most recent version of NameSieve and the study presented in this paper used a powerful mention detection[1] mechanism developed by IBM (Florian, et al., 2004). It is based on a statistical maximum-

[1] Even though we have rather loosely used the term "named entities" so far, it is more correct to say "mentions" because it includes all named, nominal, and pronominal entities. We did not make a clear distinction between named and other entities.

entropy model that recognizes 32 types of named, nominal and pronominal entities (such as PERSON, ORGANIZATION, FACILITY, LOCATION, OCCUPATION, etc), and 13 types of events (such as EVENT_VIOLENCE, EVENT_COMMUNICATION, etc). This mechanism was used to annotate every document in the TDT4 corpus[2] loaded into NameSieve. After that, we post-processed the annotation results to select the most useful entities and to reorganize them into four groups corresponding to four of the five "Ws of journalism" (*Who, Where, When, and What*) (Wikipedia, 2009), which are also frequently used in intelligence analysis. The following subsections provide a summary of the mention detection approach and the post-processing used in the presented version of NameSieve.

4.1 Mention Detection

The goal of the mention detection task is to identify and characterize the main actors in a document: the people, the locations, organizations, geo-political entities, etc. It represents one of the crucial steps in the information extraction processing pipeline, as identifying the participants in a discourse is essential to the understanding of the text: it is the first step in determining who did what where to whom. Its applications are widespread, from information extraction and template filling, to search and information retrieval, to machine translation and data mining.

Given a sentence, our goal is to identify spans of text (words) that refer to a set of pre-defined types such as persons, organizations, locations, dates, or countries. The identification of these non-overlapping and contiguous chunks of text converts into an equivalent problem of labeling each word in a sentence with a tag corresponding to the mention it belongs to (if any), as follows:

His	brother	is	John	Cairne	.
B-PER	B-PER	O	B-PER	I-PER	O

Figure 3: Mention Detection Example

- The token is not part of any mention – outside of any mention (usually O)
- The token begins a mention type X (B-X)
- The token is properly inside a mention of type X (I-X)

The B-X label type is necessary only to separate adjacent mentions of the same type, such as the case presented in Figure 3 – where several different mentions of type PERSON are directly adjacent. Such mention encoding is called the IOB representation. It is interesting to note that this mapping from token chunks to token tags is bidirectional and loss-less; one can easily go back and forth between the two representations. Historically, tagging models are preferred to the chunking type of models, mainly due to their relatively straightforward and efficient search procedure – the Viterbi dynamic-programming search, to be briefly described later. The first instance of such transformation was presented by Ramshaw and Marcus (1994), where the authors applied the IOB transformation procedure to the task of base-noun phrase chunking. Later, this method was applied to a variety of tasks, including text chunking (Ramshaw & Marcus, 1995) and named entity recognition (Tjong & Sang, 2002).

When detecting mentions, as is also true for many other natural language processing (NLP) tasks, there are many contextual, lexical and semantic clues that help in making the classification. Besides the

[2] Topic Detection and Tracking Project <http://projects.ldc.upenn.edu/TDT

obvious lexical dependency (e.g., *John* will most likely be a person, while the pronoun *we* will tend to refer to multiple people or organizations), other decision factors include: part-of-speech information, text chunking information (whether the token is part of a noun phrase, etc), whether the token appears in a predefined dictionary (is the token part of a list of names or places or organizations), how the token is labeled by other slightly different classifiers, etc. In fact, a successful mention detection system will integrate information coming from various and diverse sources; the system described here uses more than 6 streams of information. Because of our interest in using many knowledge-lean sources, we are examining those statistical systems that can easily and seamlessly integrate such information. One way to attain this goal is through the use of exponential models; in particular models trained using the maximum-entropy principle.

We are stating the sequence classification described above as the following problem: given a sequence of n words $x_{1\ldots n}$ (a sentence), find the classification sequence $y_{1\ldots n}$ which maximizes the probability $P(y_{1\ldots n} \mid x_{1\ldots n})$. This sequence probability can be computed by using the chain rule:

$$P(y_{1\ldots n} \mid x_{1\ldots n}) = \prod_{i=1} P(y_i \mid x_{1\ldots n}, y_{1\ldots i-1}) \approx \prod_{i=1} P(y_i \mid x_{1\ldots n}, y_{i-k\ldots i-1})$$

In the approximation above, we have made the regular Markov assumption that the classification y_i depends only on the last k classifications $y_{i-k\ldots i-1}$. Furthermore, to simplify notation and without any loss of generality, we can assume that the classification y_i depends only on the classification at the previous step y_{i-1} (in the more general case, one can denote the entire tuple $(y_{i-k\ldots i-1})$ by z_{i-1} and use this new variable instead).

Following the preference of allowing the modeling probability $P(y_i|x_{1\ldots n}, y_{i-1})$ to depend on multiple factors, we consider here an exponential model:

$$P(y_i \mid x_{1\ldots n}, y_{i-1}) = \frac{1}{Z(x_{1\ldots n}, y_{i-1})} \prod_j e^{\lambda_j f_j(x_{1\ldots n}, y_{i-1}, y_i)}$$

where $f_j(x_{1\ldots n}, y_{i-1}, y_i)$ are feature functions, which typically return 1 if there is some relationship between a given $x_{1\ldots n}$ and y_i (for instance, x_i is John and y_i is B-PER, or x_i is Washington and y_i is B-LOCATION) and 0 otherwise, and $Z(x_{1\ldots n}, y_i)$ is a normalizing factor that ensures that the above equation defines a proper probability. The parameters $(\lambda_j)_j$ are weights associated with the model features; higher values should be associated with the better features. These parameters can be trained using the Maximum Entropy principle – the description of the method is beyond the scope of this article, but the interested reader can read more about the training procedure (Zitouni, Luo, & Florian, 2008).

Once the probabilities $P(y_i|x_{1\ldots n}, y_{i-1})$ have been computed, one can use dynamic programming to compute the best sequence of tags by observing that

$$\max_{y_{1\ldots i}} P(y_{1\ldots i} \mid x_{1\ldots n}) = \max_{y_i} \left(\max_{y_{1\ldots i-1}} P(y_{1\ldots i-1} \mid x_{1\ldots n}) P(y_i \mid y_{1\ldots i-1}, x_{1\ldots n}) \right) \quad (1)$$

Indeed, if we use the notation $a_i(y_j) = P(y_{1\ldots i,j}|x_{1\ldots n})$, one can compute the max probability recursively as

$$
\begin{aligned}
a_0(y) &= 1 \\
a_i(y) &= \max_{y'} a_{i-1}(y') \cdot P(y \mid x_{1\ldots n}, y') \\
\max_{y_{1\ldots i}} P(y_{1\ldots i} \mid x_{1\ldots n}) &= \max_{y} a_n(y)
\end{aligned}
\qquad (2)
$$

The computation requirement for this matrix is linear in $n \cdot |Y|$, where n is the number of words in the sentence and Y is the classification space. Finding the actual sequence that maximizes Equation **(1)** is as easy as storing the individual classifications for which each of the max in Equation **(2)** happens and chaining them.

The mention detection system used in this article predicts 32 types of mentions, including persons, organizations, locations, substances, and geological objects, and 13 types of events, including business, communication, disaster, and sport events. Of these, only 9 types are used (Table 1). It also predicts for each mention whether it is a named (e.g. John Meyer), nominal (e.g. company) or pronominal (e.g. he) mention. The features used by the system to predict mention types include part-of-speech tags, text chunks (whether the word is part of noun phrase, verb phrase, prepositional phrase, etc), and whether the word is included in precompiled dictionaries of people, organizations, locations, etc. While the features themselves are language-specific, the model infrastructure is not, and models using the same framework have been built for English, Chinese, Arabic, Spanish, and Italian (Florian, et al., 2004). Table 2 shows a break-down of the performance of the English system for persons, organizations, locations and overall performance across all recognized types, as typical for the task – showing precision, recall, and their harmonic mean, the F-measure.

Table 1: Mention type distribution
(Only the most frequently used in the corpus are listed here: 7 excluded types are presented in italics)

Rank	Type	Count	Ratio	Cum. Ratio
1	**PERSON**	1 563 937	0.1818	0.1818
2	**ORGANIZATION**	1 473 310	0.1712	0.3530
3	**PEOPLE**	1 444 331	0.1679	0.5209
4	**LOCATION**	650 124	0.0756	0.5965
5	*CARDINAL*	627 096	0.0729	0.6694
6	*EVENT_COMMUNICATION*	378 423	0.0440	0.7134
7	**OCCUPATION**	357 271	0.0415	0.7549
8	**DATEREF**	333 314	0.0387	0.7936
9	**COUNTRY**	237 560	0.0276	0.8212
10	*ORDINAL*	134 763	0.0157	0.8369
11	**FACILITY**	119 783	0.0139	0.8508
12	**DATE**	110 921	0.0129	0.8637
13	*DURATION*	102 433	0.0119	0.8756
14	*EVENT_SPORTS*	100 920	0.0117	0.8873
15	*EVENT_VIOLENCE*	96 746	0.0112	0.8985
16	*EVENT_MEETING*	94 940	0.0110	0.9095

Table 2: Mention Detection Performance

Type	Precision	Recall	F-measure
Person	**92.6**	**92.6**	92.6
Organization	**70.2**	**71.0**	70.6
Location	**85.2**	**82.1**	82.6
All	76.4	77.7	77.0

4.2 Mention Processing

The mention detection mechanism presented in section 4.1 was applied to all documents in the TDT4 corpus to produce what can be called "raw annotations". Appendix 1 and 2 each show an example of the original text and the raw annotation[3] for document ZBN20001113.0400.0019, which is ranked first in the example in Figure 1. Each line under the <ENT> tag of this example represents a single entity found in the document text. It includes information such as entity type, location in the text, co-reference information, and textual representation found in the document. Out of 11 fields returned by the mention detection, NameSieve uses *Entity Type* (column 1), *Co-Reference Information* (column 8), and *Entity Text* (column 10 and 11).

The first column (Entity Type) identifies the category of each mention such as PERSON, LOCATION, and ORGANIZATION. For example, Salzburg is a "LOCATION", United Kingdom is a "COUNTRY" name, and ski_lovers are "PEOPLE". While the tagger supported numerous kinds of mention types, their distribution is not even. As Table 1 shows, 16 mention types occupied over 90% of the entity instances in TDT4 corpus. Therefore, we decided to use those 16 top entities only. Among them, 7 mention types were excluded because of low relevance for the name based browsing in NameSieve. The excluded entity types are depicted in italic font in Table 1. For example, CARDINAL types are just some casual numbers found in news articles and EVENT_COMMUNICATION types are verbs used for communications, such as "said".

We then assembled these 9 remaining entity types into four Ws (*Who, Where, When, and What*) and present them to the user tabbed browsing interface (Figure 2) so that users can work with the entities at a higher semantic level (e.g. *WHO*) and don't have to worry about minor differences among entity types (e.g. PERSON or PEOPLE).

The mapping from the entity type to the 4 W's was as follows.

Table 3: Mention type to editor's 4 W's mapping

4W	Entity Types
Who	PERSON, PEOPLE, OCCUPATION
Where	LOCATION, COUNTRY, FACILITY
When	DATEREF, DATE
What	ORGANIZATION

[3] Part of the entire annotations was listed in the appendix due to space constraints.

The mention detection mechanism also performs co-reference resolution for the identified entities, linking pronominal and nominal instances with their named antecedents (if they have one) and identifies and classifies relations between the discovered entities. This co-reference information (column 8 in Appendix 2) could be used for name disambiguation. For example, the entities "ski_lovers" and "who" (line 127 to 130) were annotated the same in the co-reference information column (ZBN20001113.0400.0019-E75). "Who" is a relative pronoun that refers to the "ski lovers" and we could see that this annotation made sense in that it referred to the same group of people. As in this example, if the tagger was able to identify the different textual expressions as identical named entities, they are given the same co-reference information in column 8. Therefore, we could use that as a unique ID for the semantically unique entities (for example, we could treat "ski_lovers" and "who" as unique entities with an identical entity ID ZBN20001113.0400.0019-E75).

In addition to this "within document" co-reference, the mention detection also supports "cross document" entity reference, which was annotated as "XDC" in the same column. For example, XDC:Cntry:United_Kingdom (line 6) can be understood as a unique entity meaning United Kingdom regardless of its form (United Kingdom, UK, or She) across all documents, because it was consistently represented as XDC:Cntry:United_Kindom in the whole corpus. Another example is the person name "Wolfgang Schussel" which appeared four times in Appendix 2 (line 148 to 151) with four different forms: "Schussel", "director", "Chancellor", and "him" (last two columns) but with the same entity representation "XDC:Per: wolfgang_schussel". We could disambiguate these three different textual representations as a unique person's name even across multiple documents, thanks to the cross document reference information. The cross document reference information here was also used as entity IDs as in the within-document co-reference information.

NE co-references within- and cross-document provided by an advanced mention detection mechanism allow NameSieve to merge various textual representations of the same NE and support browsing on a semantic level (i.e., the level of concepts meaningful for the user). Technically, document number, entity ID (co-reference information), and their frequencies are stored in the NameSieve database after the disambiguation process. Using this information, NameSieve generates its tab-based faceted browsing interface with proper NE names and counters. The only part of this process that deserves separate explanation is the selection of the best human-readable mention for each NE when presenting it to the users (e.g. "Ski Lovers" instead of "who"). While the cross document reference IDs are ready to be presented to the users after a minor heuristic-based post-processing (e.g. XDC:Cntry:Germany to Germany), the within-document entity reference IDs are not in human-readable forms at all (e.g. ZBN20001113.0400.0019-E75). We can look up the original text (column 10), such as "ski_lovers", but the problem is that they are mixed up with nouns, pronouns, relative nouns, etc., and we have no further clue to select the most optimal text representation from among them. Our first idea was simply to select the longest mention, but this method performed inconsistently. We then decided to use an external resource for this stage of disambiguation and chose Wikipedia as a dictionary. Because we are able to understand each Wikipedia entry title as a "concept", we compared every possible textual representation of a single entity with the Wikipedia titles and picked one if any of them matched one of the titles. By using this method, we were able to remove noisy textual representations of the entities that do not appear as Wikipedia entry titles.

Following is the Wikipedia-based algorithm used in the presented version of NameSieve:

(1) List the candidate entity variants from the annotated entity text (column 10).
(2) Remove stop words from the candidates.
(3) If *Where, When, What* types

a. Look up the candidates in Wikipedia titles and choose one if there is a match.
b. If not, choose the longest one among the candidates.

(4) If *Who* entity, choose the longest one among the candidates.

We did not use Wikipedia for *Who* entity types because many non-celebrity person names frequently found in news articles cannot be found in the encyclopedia (and instead could be confused with names of irrelevant celebrities); on the other hand, we expect to find place or organization names in Wikipedia.

For the *When* tab, we used a different process to normalize the entities, which are useless in the context of NE browsing (such as "today", "yesterday", "tomorrow", "this year", or "last year"). Because we had information about the release date of each news article in the TDT4 corpus, we could simply convert these relative time entities to fixed dates using simple heuristics (for example, "Today" to "Nov 12, 2000" or "This year" to "2000").

5. The Study of NameSieve

To assess the usefulness and the value of a NE-based exploratory search interface, we ran a user study of NameSieve. In our study, we wanted to assess two aspects of the approach. First, it is important to determine the usability of the approach. While potentially powerful, NameSieve's additions make the search interface more complicated and may discourage users from applying the approach. Therefore, the first group of questions we wanted the study to answer is "Will the users apply NameSieve's functionality when faced with an information exploration task?" and "Will the users appreciate NameSieve features and the whole experience of searching with an extended system?" Analysis of logs and user's subjective feedback provided the answers to these questions.

Second, it is important to know the effectiveness of the approach. Thus, we needed to answer such questions as "Will the extended system provide better ranking bringing relevant documents closer to the user attention?" and "Will the extended system help users find higher quality results and be more productive, as measured by users' selections and annotations?" These questions were harder to answer since performance evaluation requires a controlled study and an evaluation framework with a set of information exploration tasks and ground truth (i.e., information on which documents and their fragments contain content relevant to each task.)

The need to answer these two groups of questions defined our selection of the study format: a controlled user study evaluating NameSieve's impact on user performance and attitude against a baseline system without NameSieve's functionality. We used the same evaluation framework (He, et al., 2008) and the same kind of users – students in the Information Sciences who had solid search experience. Note that the selection of students as study subjects is an inherent limitation of our study: it does not allow us to generalize the findings to both professional users (such as intelligence analysts) and inexperienced users ("naïve searchers"). However, we believe that our subjects provide a good representation of non-professional, yet experienced Web searchers who (along with target users such as intelligence analysts) might also benefit from information exploration interfaces such as NameSieve.

5.1 Hypotheses and Measures

The goals of our study could be formalized as the following set of hypotheses.

H1: At the objective level, the experimental system (NameSieve) performs better.

H1-1: The users will actively use the NE features provided by the experimental system.

H1-2: The precision of the documents retrieved by the experimental system will be greater than that of the baseline system.

H1-3: The precision of the annotations made by the users using the experimental system will be greater than that of the baseline system.

H2: At the subjective level, users prefer the experimental system over the baseline system.

H2-1: Users appreciate the NE-based exploration features provided by the experimental system.

H2-2: Users are more satisfied with the experimental system.

With these hypotheses in mind, we organized a study (within subject) as a comparison between *the experimental system*, which included a full-fledged version of the NameSieve interface as presented in section 2 against *the baseline system*: a disabled version of NameSieve without filtering functionality or a NE viewer (Figure 4). This version simply performs the base search function triggered by user queries and has no support for the query reformulation, or NE and query-based filtering. This organization allowed us to not only ensure that the NE-based interface is used and appreciated by the user, but also to uncover any differences in the value of the new interface on several levels, such as system performance, user performance, and user subjective feedback.

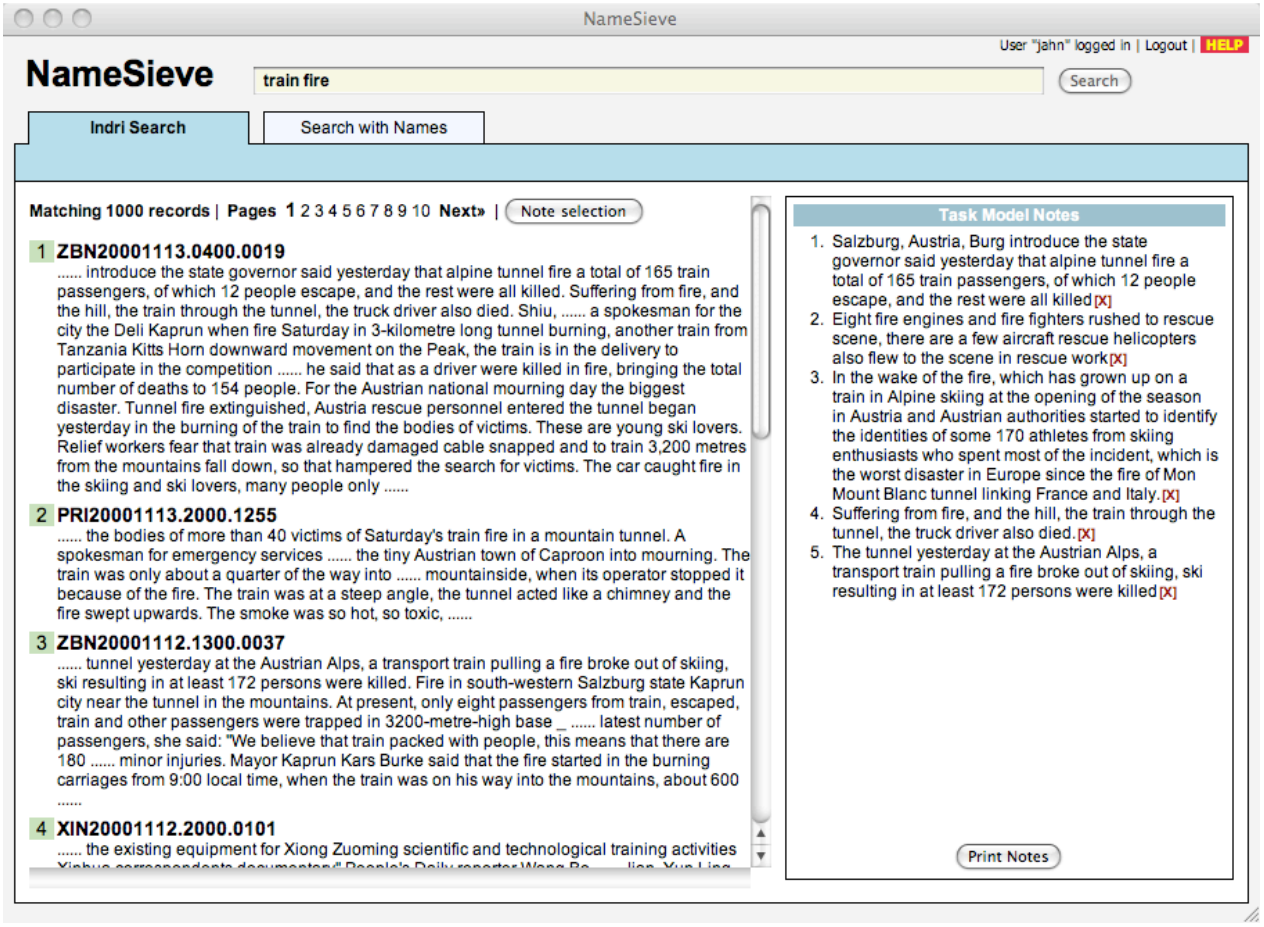

Figure 4: Baseline system without any NE-related features

We adopted several tools and measures to support the hypotheses. We mainly analyzed the log data collected during the experimental sessions in order to support the first hypothesis (H1). In order to test the hypothesis H1-1, we counted the number of times the NE features were used by the subjects. This side of the analysis helped us to understand whether the NE features were favored by the users before analyzing the real advantages provided by NameSieve. For the hypothesis H1-2, we evaluated the precision of the retrieved lists returned by the systems (system precision). By comparing this measure between the baseline and the experimental systems, we evaluated the quality of the ranked lists generated during the interaction between the users and the systems. We assume that the users can help the experimental system to generate better ranked lists by using the NE filters.

Hypothesis H1-3 addresses the precision of the notes annotated by the subjects during the experiment (user precision). The subjects were asked to select relevant passages or sentences they found and to save them to the Notebook (or a shoebox), appearing at the bottom right of Figure 1, as the final product of their activity. They could save these notes from the document surrogates of the retrieved lists or from the full-text documents. This feature was supported by both NameSieve and the baseline system. Because the notes are saved as passages, we defined a passage level precision measure and used it for comparing the precision of passages selected by the subjects. The detailed descriptions of the measures are provided in section 5.2. For the second hypotheses (H2-1 and H2-2), we used questionnaires and gathered the subjects' subjective opinions. All the evaluation results are presented from section 5.5 to 5.8.

5.2 Materials and Procedures

Our study design follows the methodology developed as a part of the task-based information exploration (TBIE) evaluation framework. The framework is constructed to examine systems in task-based information exploration (He, et al., 2008). It shares ideas with human-centered system design in the literature (Borlund, 2003; White, Kules, Drucker, & schraefel, 2006), where supporting human users in their tasks is both the focus and the criterion for examining the usefulness of the systems. The TBIE evaluation framework utilizes task scenarios that simulate the actual tasks of analysts. Under the overall umbrella of task-based information exploration, users' exploration behaviors can be categorized as first information foraging – then sense-making – for collecting useful information as part of a complex and evolving task.

The TBIE framework provides a test reference collection that was developed from the topic detection and tracking collection (TDT4). It contains 28,390 English documents and 18 task scenarios that were expanded upon from available TDT topics.

To assess retrieval effectiveness with the help of TDT4's original relevant document set, we had two human annotators go through the collection to markup passages inside each document for a given topic. The passages were annotated based on how relevant they were to the topic. In total, 1916 documents were examined with respect to their relevance. The relevance annotation produced, on average, 644.4 highly relevant passages, and 230.5 slightly relevant passages per topic. The novelty annotation produced on average 82.4 highly novel passages, and 118.3 slightly novel passages per topic. We obtained moderate inter-annotator agreement in Cohen's Kappa coefficient. We think that this is because annotations at the passage level are extremely difficult (Allan, 2003). The annotation files are independent from the source data (TDT4 collection) and can be used by anyone interested in running similar studies. In this study, we only used the relevance part of the ground truth.

The framework recommends some evaluation metrics, which includes performance-oriented measures like passage precision of selected passages, and the usability measures about the systems' support, particularly those examining the interactions between the users and the systems. Example measures include the efficiency of selecting useful information and users' subjective comments.

In this study, we adopted two measures for evaluation: system precision and user annotation precision, and they were used to test the system performance and the user performance (H1-2 and H1-3, respectively). System precision represents the ability of the system to present relevant documents in a returned ranked list during the interaction between the system and a user. Here the system precision is calculated on a ranked list at the document level. Since we are interested in how well the system pushes relevant documents to the top of the ranked list, the calculation of precision is at rank 5 and rank 10. For example, if a returned ranked list has 4 relevant documents in the top 5 and 6 relevant documents in the top 10, the precision at 5 is 4/5 = 0.8, whereas precision at 10 is 6/10 = 0.6.

The user precision was calculated at the passage level because the user's task is to select passages. Passage precision is calculated using formula (3), which is derived from a passage precision calculation (Allan, 2003). In formula (3), oll_i is the character length of the common text chunk between the snippet i and the corresponding ground truth; w_i is the weight of the ground truth combining the two annotators' mark-ups and the weight could be one of five ad hoc assigned levels: 0, 0.25, 0.5, 1, 1.25, 2; and nml_i is the character length of the part of the snippet i that has no overlap with the ground truth. Here the 0.5 associated with nml_i is the penalty weight.

$$\frac{\sum_i oll_i \times w_i}{\sum_i oll_i \times w_i + \sum_i nml_i \times 0.5} \qquad (3)$$

Two topics were selected from the 18 task scenarios that the TBIE framework provides: 40001 (Galapagos Oil Spill) and 41012 (Trouble in the Ivory Coast). As an example, the details of 40001 can be found in Appendix 3.

5.3 Data Collection

Ten subjects recruited from the University of Pittsburgh's School of Information Sciences (SIS) participated in the experiment between December 14, 2007 and January 21, 2008. To ensure that the subjects can serve as *surrogate* information analysts in the study, they were required to be native English speakers with professional training in information science and advance knowledge of information retrieval (i.e. at least a 3-credit course in the subject). Eight of the ten subjects were in graduate-level programs at SIS, while the remaining two were in the undergraduate program. Four of the ten subjects were female and the age range of all subjects was 20~36. To further recreate the information overload situation faced by professional analysts in real life, the subjects had to perform search tasks under considerable time constraints.

The experiment was conducted in one 90-minute session, consisting of a 15-minute training on the experimental and baseline systems, two 20-minute search tasks, 20 minutes for completing snippet annotation and *post-task* questionnaires, 10 minutes for breaks, and 5 minutes for a *post-session* interview. The 15-minute training included demonstrations of both versions of NameSieve (5 minutes) and a practice search task using the experimental system (10 minutes.) While the subjects were already familiar with the baseline system, they were not familiar with the features of the experimental system.

Thus, the practice task ensured that subjects had some level of proficiency with the experimental system's features before working on their two search tasks.

During each search task, subjects were given a one-page task description providing a brief background to the topic scenario and a list of questions to answer. They were instructed to search for relevant articles in the collection, analyze them, and select useful passages that provided answers to the questions in the task description. At the end of each search task, subjects annotated each snippet with the number(s) of the question(s) to which the snippet provided useful information. Subjects then completed a post-task questionnaire to assess their level of satisfaction using NameSieve for the task. Finally, after both tasks were completed, subjects filled out a brief exit questionnaire assessing their interactions with the experimental system's query and named entity filtering features versus the baseline system. The order of the two systems and the two topics were randomized among subjects to control possible learning effects.

5.4 User Activities Analysis

Before going into the main analysis and the hypothesis testing, we examined basic descriptive statistics about user activities. Table 4 and Table 5 show the average number of queries and notes made by the subjects, and compare them by the system and the topic. On average, 12.05 queries were issued and 16.15 notes were saved by the subjects. There was almost no difference between the systems in terms of the query and the note count. However, we can observe that the subjects issued a higher average number of queries with the topic 41012 than 40001 (13.6 vs. 10.5). There is a similar tendency with the number of notes made by the subjects. The subjects saved a higher average number of notes with the topic 40001 than 41012 (17.9 vs. 14.4). This data gives an interesting hint about the topic complexity. We could easily imagine that the users might have issued more queries and saved fewer notes when they were working on the more complex topic, rather than the simpler one. This tendency is repeated in the performance analysis, in sections 5.6 and 5.7.

Table 4: Average number of queries issued by the subjects

		Average Query Count	Standard Error
System	Baseline	11.6	1.63
	Experimental	12.5	2.51
Topic	40001	10.5	1.86
	41012	13.6	2.24

Table 5: Average number of notes saved by the subjects

		Average Note Count	Standard Error
System	Baseline	16.5	1.84
	Experimental	15.8	2.86
Topic	40001	17.9	8.41
	41012	14.4	2.66

5.5 Named Entity Filter Usage

The first question of our study was whether NameSieve's named entity exploration functionality was appealing enough to the subjects to be used for their exploratory searches. The answers to this question

were quite positive. While NameSieve's interface was reasonably complicated and new to all subjects, they used post-filtering 42 times in total (5 times on average among the subjects who used the NE filtering feature at least once). Among 10 users, five used the filters more than 5 times during the search sessions, three used it less than 5 times, and 2 users did not use the filters at all.

The division of the Named Entity Panel into four tabs helped us to collect usage data for each NE type. We were able to count how many times the subjects switched these tabs (Table 6) and how many entities were activated per each tab when post-filtering, which may reflect their interest in the NE and the activities to locate relevant entities. Subjects clicked on the tabs 87 times in total, i.e., over 10 times per user (except the 2 subjects who did not use the NE filtering feature). The most frequently clicked tab was *What* (31) and *Where* (24). Even though *Who* was used least frequently (10), this tab was displayed initially by default; therefore, 10 actually indicates the number of times the subjects *returned* to the *Who* tab after using some other tab. The number of entities applied during the filtering (second row) coincides with this observation. The most frequently used entities were from the *Who* tab (30) and the least from the *When* tab. This data may be understood as evidence of users' interest in the NE feature provided by NameSieve and supports the hypothesis H1-1.

Table 6: Named entity tab switch

NE type	Who	Where	When	What	Total
Tab switching frequency	10	24	22	31	87
Number of entities used for post-filtering per tab	30	23	9	28	90

5.6 System Performance Analysis

The second question of this study is whether a search system equipped with named entity exploration functionality could better support users in finding relevant information (H1-2). For search systems, system performance is traditionally assessed by its ability to place relevant documents high in the ranked list of search results. Thus, to compare the performances of experimental and baseline systems we need to consider ranked lists of results obtained by the user when working on the same exploration task in both systems and check which system is better able to "push" task-relevant documents to the top of the lists of results. This ability is typically measured as precision at rank 5 and 10, i.e., proportion of relevant results among the top 5 and top 10 documents. As discussed in section 5.2, we have ground truth information on the topics used in the experiments and were able to easily calculate the task-level precision of each ranked document list returned by the systems. Note, however, that a comparison of the NameSieve interface with a baseline interface is not as straightforward as a comparison between two regular search systems. The problem is that NameSieve changes the nature of the retrieval and ranking process, turning it from a traditional one-stage to a two-stage process. In the first stage, the user issues a regular query and observes the list of results and extracted NEs. In the second stage, the user selects one or more NE to post-filter and re-rank the original set of results. Thus, to examine the effect of NameSieve, we need to distinguish between ranked lists obtained in the first stage (before post-filtering) and at the second stage (after engaging at least one NE filter). While we expected that NameSieve would deliver better results after post-filtering, it is hard to expect that its performance on the first stage would

be better than the performance of a baseline system since the query formulation and search stage in NameSieve does not differ from baseline system. Moreover, we might expect that the first stage performance of NameSieve would be worse, since the users – enabled with powerful post-filtering – would become less careful when formulating the original query. Therefore, we separated calculated NameSieve performance for first-stage lists (experimental without NE filters engaged) from the second-stage lists generated using the experimental system with the NE filters.

Figure 5 shows the comparison of the system performance in terms of document level precision at rank 5 and 10 during the experiment. The results confirm our expectations. The experimental system with engaged NE filters (rightmost column) demonstrated nearly perfect performance with average precision 1.0 at rank 5 and 0.99 at rank 10. This precision was significantly higher than the average precision achieved without post-filtering by both the baseline system and the experimental system without post-filters engaged (Wilcoxon rank sum test, $p < 0.01$ and $p = 0.02$ for rank 5 and 10, respectively). It supports our hypothesis that the use of NameSieve's visualization and post-filtering interface significantly improves system performance for information exploration tasks. As we expected, we also observed a slight decline in the experimental system's precision on stage 1 in comparison with the baseline system. While this difference appeared to be insignificant, it could hint that users gradually become "less careful" with their first-stage query formulation in NameSieve. While we found no formal evidence in favor of this hypothesis, we think that this issue needs further exploration.

The differences in the experimental system's performance with and without NE filters are reminders that the mere presence of new features in a system does not automatically make the system more efficient. The user needs to actively implement the advanced features to obtain better system performance. The voluntary nature of the NE interface (it is left to the user to use filters or not) may limit the impact of the system in a practical context, since system performance for the users who choose not to use the NE interface (2 out of 10 users in our study) will hardly be improved.

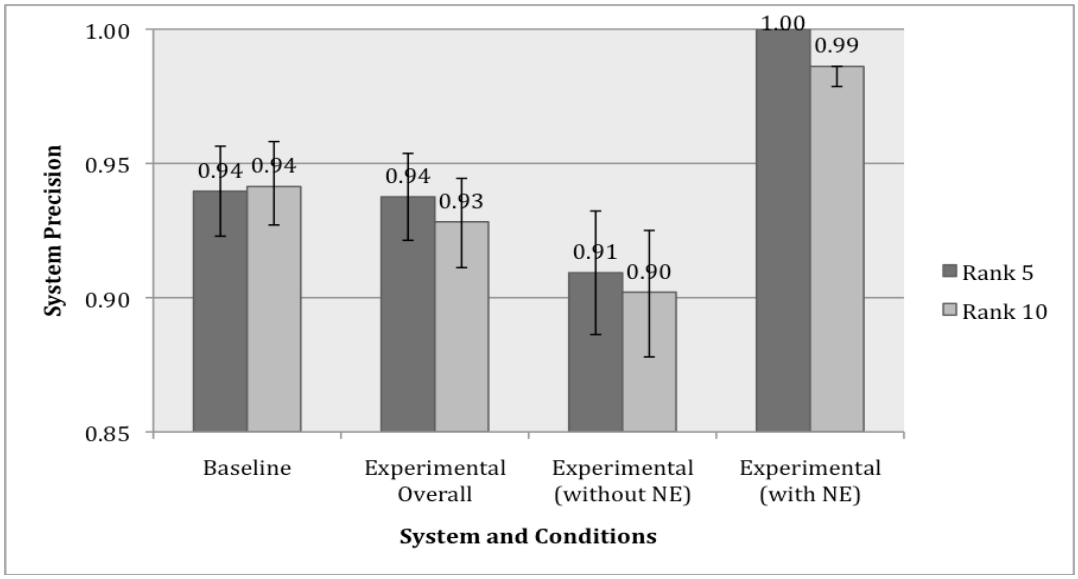

Figure 5: System performance comparisons

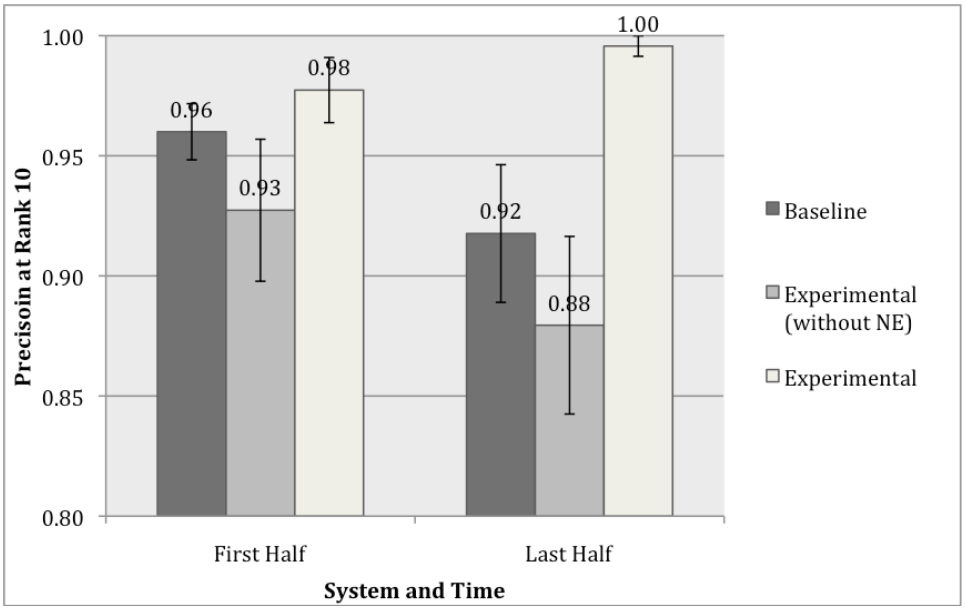

Figure 6: System performance over time (rank 10)

We also analyzed how the performance changed during the interaction between the subjects and the system. Each session lasted 20 minutes, and we could record the changes in the system's performance as the session progressed. Figure 6 shows the changes in the system's performance (precision at rank 10) between the first and last half of the sessions (10 minutes each). Both the baseline and the experimental (without NE) systems (black and gray columns) show that performance decreased in the last half of the session (the likely cause is that it was getting incrementally harder to discover new relevant documents using queries), while the experimental system's performance (white columns) improved (probably as a result of uses gaining experience in using NameSieve). As discussed before, the experimental system without using NE filters behaved identically to the baseline and the pattern of the performance change was the same as that of the baseline system. The experimental system (white columns) with NE filters showed improved performance over the baseline (black and gray columns); the difference was statistically significant in the last half of the session (Wilcoxon rank sum test, $p = 0.01$).

5.7 User Performance Analysis

In addition to the system level performance analysis, we were also able to calculate the precision of the users' annotations during the experimental sessions and to test our hypothesis H1-3. The passage precision score here was calculated against the ground truth, and the detail was described in section 5.2. Figure 7 compares user performances by topic (40001 and 41012) and time (first and last half of the session). The tendency revealed in the previous analysis is not clearly shown in this graph. Even though the overall precision in user annotations may look slightly better for the experimental system than the baseline, the difference was not statistically significant. However, when we examined this statistic separately by topic, we found that subjects working on the more difficult topic, 41012 (Table 7), created better annotations with the experimental system (0.75 vs. 1.0) during their 0~10 minute periods (first half session), and this difference was statistically significant (Exact Wilcoxon rank sum test, $p = 0.01$). While the average data make it appear that they performed slightly worse using the experimental system for the simpler topic, 40001, this difference was not statistically significant.

Table 7: System performance comparison by topic

	40001	41012
Rank 5	0.97	0.88
Rank 10	0.96	0.87

This fact was encouraging because the subjects showed significantly improved performance with the more complex topic. In addition, they were able to create better annotations in their first half sessions, which means their annotating behavior was very efficient. The reason why the overall user performance using the experimental system was not as good as the system's performance might be described as follows. The system supports NE filter-based exploration and provides improved retrieval lists as revealed by the system performance analysis. However, the document surrogates and the news text provided by NameSieve from which users were asked to make annotations were not different from those provided by the baseline. As the previous section shows, the current version of NameSieve can increase the user's ability to locate relevant documents, but provides no advantage in comparison with the baseline in locating the right fragments inside these documents. Therefore, the difference in user performance between the two systems may have decreased despite the initial support of the improved ranked lists provided by NameSieve.

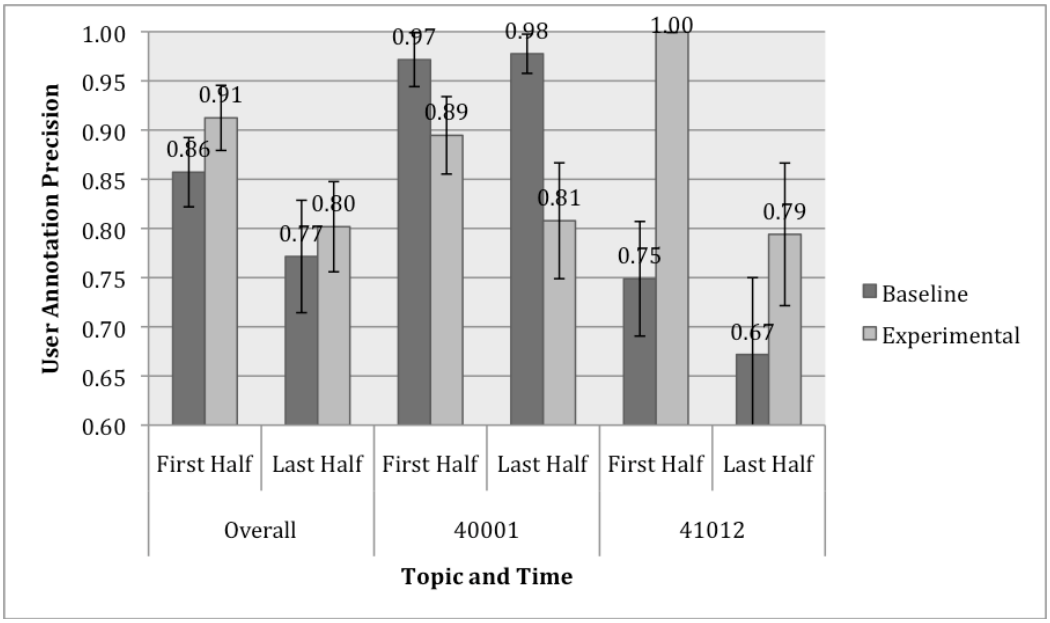

Figure 7: User performance

5.8 User Feedback Analysis

Following each search task, subjects were given a post-questionnaire to assess their satisfaction with the version of NameSieve assigned to them for that task. For all questions, subjects were asked to rate their level of agreement from 1 (Not at All) to 5 (Extremely). At the end of the questionnaire, subjects were given the opportunity to write any additional comments they had about NameSieve or the preceding search task, in general.

For the experimental version only, subjects were asked to rate the utility of the features related to filtering and NE viewing: the ability to filter search results by query terms; display of the document

counts for each query term; displaying the NEs; the ability to filter search results by NEs; separation of NEs into groups, i.e. *Who*, *What, Where, When*; using larger fonts to display higher-ranked NEs; relevancy of NEs to results of queries; and display of the ranks of NEs via scroll-over pop-up text. For both systems, subjects were asked to rate their familiarity with the assigned topic; the sufficiency of news provided; utility of the document summaries in the search results; their ability to find useful passages; the system's ease of use; and overall satisfaction with the system. After both search tasks were completed, subjects filled out an exit questionnaire asking them to compare the utility of the experimental version's filtering tools versus the baseline system, and were interviewed for 5-10 minutes to further explain their impressions of NameSieve.

Based upon their questionnaire responses (Table 8) as well as oral and written comments, subjects had an overall positive opinion of the experimental system's NE features. Six of the ten subjects noted that larger font sizes for higher-ranked named entities, grouping named entities by *Who/What/When/Where*, and query-term filtering were all very helpful in locating important information. Four of the subjects also noted that simply viewing the NEs gave them a better sense of the unfamiliar topic assigned to them, helping them to construct queries that yielded relevant snippets. Three subjects also liked the highlighting of query terms in the result snippets, and asked that the same highlighting be applied to full articles.

Three of the five subjects who completed their first search task on the experimental system also said that they wished that NE features had been available to them for their second task. This feedback helps to explain the mean post-task questionnaire responses to questions pertaining to the utility of the NE features shown in Table 8. While subjects' question responses were not as positive as their oral and written comments, there was a noticeable pattern in responses by subjects assigned to the experimental system for the first task (Sequence 1) versus those for their second task (Sequence 2). Table 8 shows that subjects who tried the experimental system *after* the baseline (i.e., Sequence 2), expressed much more positive opinions about the NE faceted browsing interface main features, especially for its key functionality, NE filtering. Although Chi-squared tests indicated the differences in responses were not significant, it hints that the subjects were not able to fully appreciate faceted browsing with NE until they could compare their experiences solving realistic tasks *with* and *without* this interface.

Table 8: Mean post-task questionnaire responses by sequence to questions applicable to the Experimental System ONLY.

Question	Seq. 1 (n=5)	Seq. 2 (n=5)	Overall (n=10)
Utility of Query-Term Filtering	3.4	3.4	3.4
Utility of Document Counts	2.4	3.2	2.8
Utility of Displaying NEs	3.2	3.8	3.5
Utility of NE Filtering	2.8	4.0	3.4
Utility of NE Grouping	3.4	3.4	3.4
Utility of NE Font Sizes	3.8	4.6	4.2
Accuracy of NEs	3.8	3.4	3.6
Utility of Pop-up Text (NE Ranks)	2.8	2.0	2.4

Table 9: Mean post-task questionnaire responses (by system) to questions applicable to both systems.

Question	Base. (n=10)	Exp. (n=10)
Familiarity with Topic	1.5	1.1
Sufficiency of News	4.4	3.9
Utility of Document Summaries	3.7	3.6
Ability to Find Useful Passages	3.6	4.0
Ease of Use	3.9	3.8
Overall Satisfaction	3.9	3.7

Chi-square tests were performed on the questionnaire data to determine if there were any significant differences in subject responses between the two versions of NameSieve.

Table 9 shows the mean post-questionnaire responses by system to questions applicable to both systems. While there were no significant differences between users' subjective ratings of the baseline and experimental systems, the consistently positive ratings for the experimental system suggest that NE features were helpful additions to the baseline search system, despite the features' relative novelty and the sequence effect shown in Table 8.

Table 10: Mean exit questionnaire responses, overall and grouped by system sequence.

Q#	Question	Seq. 1 (n=5)	Seq. 2 (n=5)	Overall (n=10)
1	Does the ability to view named entities extracted from search results provide better support for finding useful information compared to traditional search?	3.8	3.6	3.7
2	Is the ability to use named entities to further filter search results an important addition to traditional search?	3.8	3.8	3.8
3	Is the ability to filter the search results containing specific query terms an important addition to traditional search?	4.4	4.6	4.5

This sequence effect is also confirmed by answers to the exit questionnaire, which was administered after the subjects had worked with both systems. The assessments of the NE features in this questionnaire (Table 10) were higher, on average, than those in the post-task questionnaire about the experimental system (Table 8), mainly because Sequence 1 subjects provided a much more positive opinion about NE features after trying to work with the second topic without them. In addition, we can

hypothesize that more experience in working with the kinds of tasks we used in the study helped the subject to better appreciate more advanced NE features. This indicates that better training should have been provided on both the baseline and experimental systems, so subjects could become familiar with the capabilities of both prior to the start of the experiment. A longer user study with multiple tasks on each system could also have allowed subjects to use the NE features more often and to appreciate them more.

6. Related Studies

By the nature of the application area and the key technology used in NameSieve, the system belongs to two intersecting areas: information systems for intelligence analysis and exploratory search systems. This section provides a brief review of the most similar works in these two areas.

Intelligence Analysis is one of the most challenging human information processing activities and requires an analyst to provide extensive information support efficiently in a given context under time constraints. A range of information systems have been developed to support intelligence analysts over the last decade. The best framework for understanding and comparing a multitude of existing systems was suggested by Pirolli and Card (2005) and its use for a systematic review of tools for intelligence analysis was demonstrated by Card (2007). The Pirolli-Card framework recognizes two major stages in the work of intelligence analysis – information foraging and sense-making – and several smaller overlapping subprocesses. The goal of the foraging stage is to assemble a rough collection of resources focusing on recall rather than precision (not to miss important things). The goal of the sense-making stage is to "make sense" of the collected information; extracting facts, regularities, and forming ideas and theories. Since early foundational work on sense-making (Russell, Stefik, Pirolli, & Card, 1993) and information foraging (Pirolli & Card, 1999), a range of projects have specifically focused on these two stages. In the context of the Pirolli-Card framework, NameSieve can be classified as an information foraging tool, which uses information visualization to help analysts assemble this rough collection (sometimes called a "shoebox"). We can name a few similar tools that use information visualization and were specifically developed for intelligence analysis (Card, 2007; Luo, Fan, Yang, Ribarsky, & Satoh, 2006; Proulx, et al., 2006). However, the majority of tools in this category are not analysis-specific (see a brief review below). NameSieve is different from all these tools in its NE-based approach to visualize and explore a set of documents. While the use of NEs in intelligence analysis has been explored by a few other projects (Bier, Card, & Bodnar, 2008; Gersh, Lewis, Montemayor, Piatko, & Turner, 2006), these projects use NEs for sense-making, while NameSieve uses it for information foraging.

As an exploratory search tool, NameSieve combines the ideas of two research streams. One aspect of NameSieve is the visualization of the conceptual content of a retrieved set of documents. In this aspect, it is similar to other systems that attempt to visualize search results based on keyword-level content, including such classic systems as Tilebars (Hearst, 1995) and VIBE (Olsen, Korfhage, Sochats, Spring, & Williams, 1993). It was also informed by research on clustering and organization of retrieved results by their semantic similarity (Chen & Dumais, 2000; Leuski & Allan, 2004). Our system can be considered as an expansion of the idea of result clustering allowing multiple clustering by NE. Each NE serves as a cluster label and can instantly call up a cluster of documents related to this NE.

The idea of extracting and visualizing NEs in the retrieved set of documents is an extension of our own work on making keyword-level user models visible (Ahn, Brusilovsky, Grady, He, & Syn, 2007). For NE visualization, we used the same format, which was influenced by the modern approach to present tag clouds in social tagging systems. While working on this project, we discovered a few other approaches

driven by the same idea: extracting and visualizing information from the list of search results. Kuo, Hentrich, Good, and Wilkinson (2007) suggested extracting keywords from the returned documents and presenting it in the form of a tag cloud. WordBars3 system (Hoeber, 2007; Hoeber & Yang, 2008) extracts the top 20 keywords from retrieved snippets and allows the user to specify the importance of these keywords and re-filter the results. The project presented in this paper differs from the works mentioned above in several aspects: the breadth and depth of information extraction, the opportunities to use the extracted information for interactive exploration of the results, and - most importantly - our attempt to move from keyword representation to the semantic level by using NEs.

Another aspect of NameSieve is the use of recognition-based browsing (rather than recall-based search) to allow users to explore a collection of documents. The idea of repeated narrowing of the filtering of retrieved documents by clicking on extracted NEs was inspired to some extent by the modern stream of work on faceted interfaces. In that sense, extracted NEs can be considered as replacements for facet labels when no metadata is available. Since the early work on faceted search (Yee, Swearingen, Li, & Hearst, 2003), faceted interfaces such as Relation Browser (Capra & Marchionini, 2008), mSpace (Wilson & schraefel, 2008) and faceted web search (Kules & Shneiderman, 2008) were recognized and explored as effective tools for exploratory search. Within the area of faceted interfaces, NameSieve belongs to a small group of systems that attempts to build a faceted interface on the fly rather than using an existing classification (Dakka & Ipeirotis, 2008; Dash, Rao, Megiddo, Ailamaki, & Lohman, 2008). In this group, NameSieve is distinguished by its use of NE categories for facet organization.

7. Conclusions and Future Work

In this paper, we presented a non-traditional approach to building a better information access system using named entities, a popular type of semantic annotation. The proposed approach was implemented in the NameSieve system, which attempted to support the information exploration work of an intelligence analyst. NameSieve transparently presents a summary of search results in the form of an NE "cloud," while allowing the analyst to further explore the results using this cloud as a faceted browsing interface. The goal of NameSieve was to help the user in sense-making, query formulation, and manipulating search results. Our study demonstrated that we achieved some of our goals. The new interface was actively used and positively evaluated by the subjects. It enabled them to bring most relevant documents closer to the surface and achieve better performance working with a more difficult topic.

While the study provides some strong support in favor of NE-based exploratory search, its findings should not be generalized beyond its limitations. Most importantly, our study used surrogate intelligence analysts as subjects. While we make all attempts to recruit users as close to the target users as possible and to place the users in a comparable information overload context, we cannot make any claims about professional intelligence analysts' performance or attitude to the presented interface. In addition, the size of our study (10 subjects) is not sufficient to make reliable claims about all benefits of the NE interface. Another limitation of the study is that subjects had a relatively short time to master a relatively sophisticated NE-based exploratory search interface. This was our concern before the start of the study and we attempted to address it by providing some training to help the subjects to familiarize themselves with the new features. Yet, the apparently more positive user feedback about NameSieve when used for the second task hints that users need more experience with both the system and the kind of tasks used in the study to fully appreciate and exploit the innovative interface. While we can speculate that the benefits of the NE interface will increase as users gain more experience using it, a much longer user study (which we are planning to perform in the future) is required to state this reliably.

Finally, the search task performed by the users in our study cannot be considered a fully exploratory search task due to the presence of reasonably clearly-defined questions to answer. Evaluating an exploratory search interface using this kind of task limited our ability to explore the true value of this interface. However, as we explained above, the choice of task was motivated by the presence of detailed ground truth data, which allowed us to compare NameSieve's interface with a traditional search interface on a fine-grained, reliable basis. We consider this choice a compromise between realism and experimental control. For a good discussion of the tasks that can be used to properly evaluate exploratory search interfaces, we refer the reader to Kules and Capra (2008).

While the focus of this paper is the innovative interface for NE-based exploratory search, we want to stress that the use of both an advanced mention detection mechanism (one that is able to detect and co-reference multiple mentions of the same entity) and the user-oriented post-processing were as critical to the success of the NameSieve interface as the interface itself. As we mentioned above, an earlier experiment with the NameSieve interface built upon a simpler NE extractor without co-referencing failed to show the benefit of NE browsing. It demonstrated that unresolved NEs are more confusing than helpful to users.

The NameSieve interface presented in this paper was based on a specific kind of semantic annotation; however, we believe that similar information exploration interfaces could be built for users of other kinds of annotations, such as ontological concepts. In our future work, we hope to explore this opportunity as well.

In a broader context, it is important to observe again that in this work we switched from AI to HCI techniques to provide improved support for information exploration tasks. However, our long-term goal is to combine AI and HCI approaches to get "the best of both worlds". Pioneering work of other teams (Gotz, Zhou, & Aggarwal, 2006) show the promise of this direction. In our future work, we intend to combine the ideas of user-controlled personalized search explored earlier (Ahn, et al., 2007; Ahn, et al., 2008) with NE-based information exploration. Personalization should extend the power of an NE-based exploration interface. In turn, this interface could extend the bandwidth of user modeling, enabling us to maintain better knowledge and interest models of the users.

References

Ahn, J.-w., Brusilovsky, P., Grady, J., He, D., & Syn, S. Y. (2007, May 8-12, 2007). *Open user profiles for adaptive news systems: help or harm?* Paper presented at the the 16th international conference on World Wide Web, WWW '07, Banff, Canada.

Ahn, J.-w., Brusilovsky, P., He, D., Grady, J., & Li, Q. (2008, April 21-25, 2008). *Personalized Web Exploration with Task Models.* Paper presented at the the 17th international conference on World Wide Web, WWW '08, Beijing, China.

Allan, J. (2003). *HARD Track Overview in TREC 2003 High Accuracy Retrieval from Documents.* Paper presented at the The Twelfth Text Retrieval Conference.

Bier, E. A., Card, S. K., & Bodnar, J. W. (2008). *Entity-Based Collaboration Tools for Intelligence Analysis.* Paper presented at the IEEE Symposium on Visual Analytics Science and Technology, VAST 2008, Columbus, Ohio.

Borlund, P. (2003). The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research, 8*(3).

Capra, R., & Marchionini, G. (2008). *The relation browser tool for faceted exploratory search.* Paper presented at the JCDL '08: Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries, Pittsburgh PA, PA, USA.

Card, S. K. (2007, January 31 - February 4, 2007). *Leverage Points and Tools for Aiding Intelligence Analysts.* Paper presented at the The HCIC 2007 Winter Workshop, Snow Mountain Ranch, Fraser, Colorado.

Chen, H., & Dumais, S. (2000, April 2000). *Bringing order to the web: Automatically categorizing search results.* Paper presented at the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI'2000), The Hague, The Netherlands.

Dakka, W., & Ipeirotis, P. G. (2008, 7-12 April 2008). *Automatic Extraction of Useful Facet Hierarchies from Text Databases.* Paper presented at the IEEE 24th International Conference on Data Engineering, Cancun, Mexico.

Dash, D., Rao, J., Megiddo, N., Ailamaki, A., & Lohman, G. (2008, October 26-30, 2008). *Dynamic Faceted Search for Discovery-driven Analysis.* Paper presented at the The 17th ACM conference on Conference on information and knowledge management: CIKM '08, Napa Valley, CA, USA.

Demner-Fushman, D., & Oard, D. W. (2003). *The Effect of Bilingual Term List Size on Dictionary-Based Cross-Language Information Retrieval.* Paper presented at the 36th Annual Hawaii International Conference on System Sciences (HICSS'03) - Track 4, Hawaii.

Florian, R., Hassan, H., Jing, H., Kambhatla, N., Luo, X., Nicolov, N., et al. (2004). *A Statistical Model for Multilingual Entity Detection and Tracking.* Paper presented at the Human Language Technologies 2004: the Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL'04).

Gersh, J., Lewis, B., Montemayor, J., Piatko, C., & Turner, R. (2006). Supporting insight-based information exploration in intelligence analysis. *Communications of the ACM, 49*(4), 63-68.

Gotz, D., Zhou, M. X., & Aggarwal, V. (2006, October 31 - November 2, 2006). *Interactive Visual Synthesis of Analytic Knowledge.* Paper presented at the IEEE Symposium on Visual Analytics Science and Technology, VAST 2006, Baltimore, MD.

He, D., Brusilovsky, P., Ahn, J.-w., Grady, J., Farzan, R., Peng, Y., et al. (2008). An evaluation of adaptive filtering in the context of realistic task-based information exploration. *Information Processing and Management, 44*, 511-533.

Hearst, M. A. (1995, May 7-11, 1995). *TileBars: Visualization of term distribution information in full text information access.* Paper presented at the CHI'95, Denver.

Hoeber, O. (2007, November 2-5, 2007). *Exploring Web search results by visually specifying utility function.* Paper presented at the International Conference on Web Intelligence, Silicon Valey, CA, USA.

Hoeber, O., & Yang, X. D. (2008). Evaluating WordBars in exploratory Web search scenarios. *Information Processing & Management, 44*(2), 485-510.

Khalid, M. A., Jijkoun, V., & Rijke, M. d. (2008). *The Impact of Named Entity Normalization on Information Retrieval for Question Answering.* Paper presented at the European Conference on Information Retrieval ECIR 2008.

Kules, B., & Capra, R. (2008, October 23, 2008). *Constructing Exploratory Tasks for a Faceted Search Interface.* Paper presented at the the Second Workshop on Human-Computer Interaction and Information Retrieval, Redmond, Washington, USA.

Kules, B., & Shneiderman, B. (2008). Users can change their web search tactics: Design guidelines for categorized overviews. *Information Processing & Management, 44*(2), 463-484.

Kumaran, G., & Allan, J. (2004). *Text Classification and Named Entities for New Event Detection* Paper presented at the the 27th annual international ACM SIGIR conference on Research and development in information retrieval.

Kuo, B. Y.-L., Hentrich, T., Good, B. M., & Wilkinson, M. D. (2007, May 8-12, 2007). *Tag Clouds for Summarizing Web Search Results.* Paper presented at the the 16th international conference on World Wide Web, WWW '07, Banff, Canada.

Leuski, A., & Allan, J. (2004). Interactive information retrieval using clustering and spatial proximity. *User Modeling and User Adapted Interaction, 14*(2-3), 259-288.

Luo, H., Fan, J., Yang, J., Ribarsky, W., & Satoh, S. i. (2006, October 31 - November 2, 2006). *Exploring Large-Scale Video News via Interactive Visualization.* Paper presented at the IEEE Symposium on Visual Analytics Science and Technology, VAST 2006, Baltimore, MD.

Mandl, T., & Womser-Hacker, C. (2005). *The Effect of Named Entities on Effectiveness in Cross-Language Information Retrieval Evaluation.* Paper presented at the ACM SAC'05, Santa Fe, NM.

Marchionini, G. (2006). Exploratory search: From finding to understanding. *Communications of the ACM, 49*(4), 41-46.

Micarelli, A., Gasparetti, F., Sciarrone, F., & Gauch, S. (2007). Personalized search on the World Wide Web. In P. Brusilovsky, A. Kobsa & W. Neidl (Eds.), *The Adaptive Web: Methods and Strategies of Web Personalization* (Vol. 4321, pp. 195-230). Berlin Heidelberg New York: Springer-Verlag.

Mihalcea, R., & Moldovan, D. I. (2001). Document Indexing Using Named Entities *Studies in Informatics and Control*.

Oard, D. W. (2002). *When You Come to a Fork in the Road, Take It!* Paper presented at the Proceedings of SIGIR2002 workshop "Cross-Language Information Retrieval: A Research Roadmap".

Olsen, K. A., Korfhage, R. R., Sochats, K. M., Spring, M. B., & Williams, J. G. (1993). Visualisation of a document collection: The VIBE system. *Information Processing and Management, 29*(1).

Pablo-Sanchez, C. d., Martınez-Fernandez, J. L., & Martınez, P. (2005). *Named Entity Processing for Cross-lingual and Multilingual IR applications.* Paper presented at the CLEF 2005.

Petkova, D., & Croft, W. B. (2007). *Proximity-based Document Representation for Named Entity Retrieval.* Paper presented at the CIKM07.

Pirolli, P., & Card, S. (1999). Information foraging. *Psychological Review, 106*(4), 643-675.

Pirolli, P., & Card, S. K. (2005, 2-4 May 2005). *The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis.* Paper presented at the 2005 International Conference on Intelligence Analysis, McLean, VA.

Pizzato, L. A., Molla, D., & Paris, C. (2006). *Pseudo Relevance Feedback Using Named Entities for Question Answering.* Paper presented at the Australasian Language Technology Workshop 2006.

Proulx, P., Tandon, S., Bodnar, A., Schroh, D., Harper, R., & Wright, W. (2006, October 31 - November 2, 2006). *Avian Flu Case Study with nSpace and GeoTime.* Paper presented at the IEEE Symposium on Visual Analytics Science and Technology, VAST 2006, Baltimore, MD.

Ramshaw, L., & Marcus, M. (1994). *Exploring the Statistical Derivation of Transformational Rule Sequences for Part-of-Speech Tagging.* Paper presented at the The Balancing Act: Proceedings of the ACL Workshop on Combining Symbolic and Statistical Approaches to Language.

Ramshaw, L., & Marcus, M. (1995). *Text Chunking Using Transformation-Based Learning.* Paper presented at the Proceedings of the Third Workshop on Very Large Corpora.

Russell, D., Stefik, M., Pirolli, P., & Card, S. (1993). *The cost structure of sensemaking.* Paper presented at the the SIGCHI conference on Human factors in computing systems, CHI '93.

Tjong, E., & Sang, K. (2002). *Introduction to the CoNLL-2002 shared task: language-independent named entity recognition.* Paper presented at the COLING-02: proceeding of the 6th conference on Natural language learning.

White, R. W., Kules, B., Drucker, S. M., & schraefel, m. c. (2006). Supporting exploratory search. *Communications of the ACM, 49*(4), 37-39.

Wikipedia (2009). Five Ws. *Wikipedia*, 2009, from http://en.wikipedia.org/wiki/Five_Ws

Wilson, M. L., & schraefel, m. c. (2008, June 16-20, 2008). *A Longitudinal Study of Exploratory and Keyword Search.* Paper presented at the Joint Conference on Digital Libraries, JCDL 2008, Pittsburgh, Pennsylvania, USA.

Wu, D., He, D., Ji, H., & Grishman, R. (2008). *The Effects of High Quality Translations of Named Entities in Cross-Language Information Exploration.* Paper presented at the The 2008 IEEE International Conference on Natural Language Processing and Knowledge Engineering, Beijing, China.

Yang, Y., Zhang, J., Carbonell, J., & Jin, C. (2002). *Topic-conditioned novelty detection*. Paper presented at the Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining.

Yee, K.-P., Swearingen, K., Li, K., & Hearst, M. (2003, April 5-10, 2003). *Faceted metadata for image search and browsing.* Paper presented at the ACM Conference on Human Factors in Computing Systems, CHI 2003, Ft. Lauderdale, FL.

Zitouni, I., Luo, X., & Florian, R. (2008). A Statistical Model for Arabic Mention Detection and Chaining. In A. Farghaly (Ed.), *Arabic Computational Linguistics*: Center for the Study of Language and Information.