Using Category Information for Relationship Exploration in Textual Data

Yan Qu

School of Information, University of Michigan, 1075 Beal Ave., Ann Arbor, MI 48105 yqu@umich.edu

George Furnas

School of Information, University of Michigan, 1075 Beal Ave., Ann Arbor, MI 48105 furnas@umich.edu

Ben Walstrum

School of Information, University of Michigan, 1075 Beal Ave., Ann Arbor, MI 48105 walstrum@umich.edu

In the comprehension of textual data, it is critical for people to perceive relationships between topics. This work explores two approaches that use text categorizations to reveal underlying relationships: the Overlap approach, which visualizes overlaps between categories, and the Search approach, which shows topical search results in the context of categories. The effectiveness of these approaches is tested using various types of relationship questions. Our results show that the Overlap approach improves users' performances in relationship exploration tasks. Conversely, the Search approach did not show the same effectiveness, primarily due to the Vocabulary Problem. Design implications are drawn from the experiment.

Introduction

Comprehending large volumes of textual information is an increasingly common task for knowledge workers, from intelligence analysts to purchasing agents. It requires an understanding of basic topics referred to in the texts as well as the relationships between those topics. For example, comprehending a set of news stories requires understanding not only the "who, what, where, and when" of the stories (the people, events, countries, etc.), but also understanding their interrelationships (e.g. Did a coup in a country affect the later election? Did any other countries get involved in the coup?) In this paper, we explore how categories built upon textual information can help people

in exploring such relationships.

Categorization is often an essential step when attempting to make sense of large quantities of textual information. Manual categorizations have a long history, ranging from library classifications (Dewey, LOC), to subject classification systems used in medicine and science, to more recent social tagging mechanisms. In recent decades many computer-based approaches have been explored including automatic clustering and classifications of texts. In both cases, such categorizations enable people to quickly skim through the data and understand the main topics in it.

Categories should be similarly useful in understanding relationships between the topics. By laying out the basic topics of the data set, categories provide a useful and often visualizable infrastructure for further relational analysis. In this paper, we explore two distinct approaches - the Overlap approach and the Search approach - that use category information to reveal collocations of topics, i.e., documents which contain multiple topics and hence likely discuss relationships between them.

The Overlap Approach

Since the categories represent different topics in a set of texts, the overlap or intersection between different categories comprises documents mentioning multiple topics. Displaying such overlaps makes available texts particularly be helpful in relationship exploration tasks. For example, any news articles detailing both a coup attempt and a subsequent election might provide useful information on how these two events were related.

The overlaps between categories are common in multi-faceted categorization, where each facet is used to represent the data from a different perspective. For example, a set of news articles can be organized into several categorical schemas based on facets such as event-type, geo-political region, or chronology. The overlap of categories from different categorical schemas reflects relationships between separate facets of the data. For example, the overlaps between an event category and a country-category might provide information about the involvement of the corresponding country in the corresponding event.

Even within the same categorical schema, there may be overlaps between categories. For example, when we organize news articles into categories representing different events, one article can be assigned to several categories if it mentions multiple events. The overlaps then suggest there may be relationships between the different events.

The Search Approach

As an alternative, consider a hybrid system that has both pre-categorization outlining the

main topics in a collection of texts, and full text search. The results of a search, when understood in the context of the categorization, reveals documents mentioning both the query topic and the categorical topics. Such documents may provide useful information on the relationship between the corresponding topics. For example, a query detailing a news event viewed in the context of the country-based categorical schema can inform about the involvement of different countries in this news event. Therefore, enabling people to see the search results in the context of the categories can be beneficial in relationship exploration tasks.

We are motivated to explore the hybrid Search approach for several reasons: 1) Insofar as many information systems still enforce *exclusive* categorizations on textual data (either by human-drawn criteria or machine learning algorithms), search provides a suitable solution for discovering the collocation of topics in those systems. 2) Users can choose the search topic freely in their relationship exploration task, allowing them to test various hypotheses without being restricted to existing categories. 3) Search is a popular tool in current information systems. Therefore it has a low learning cost for users.

The paper proceeds as follows: First we review the literature on related techniques and systems that help people explore relationships in textual data. Second, we introduce a system supporting the Overlap and Search approaches. Third, we describe the evaluation of the effectiveness of the two approaches for helping people explore relationships in a data set with multi-faceted, non-exclusive categorical schemas. Finally, we discuss interesting issues raised in the experiment and future work.

Related Work

In this section, two bodies of related work are surveyed briefly. First we address works on automatically detecting relationships in textual data. Second, we discuss various existing visualization techniques that can be used in revealing relationships in categorized text data.

Detecting Relationships in Textual Data

Researchers in the Natural Language Processing and Text Mining fields have long been interested in relationships within textual data. Linguistic models and machine learning techniques are used in automatically detecting relations, patterns, and structures in textual data at various granularities. At the word level, relationships among lexical items can be detected by using grammatical knowledge and statistical methods on large text corpora (Hindle, 1990; Hearst 1998). Moving up to the sentence/discourse level, based on theories of rhetorical and discourse structures (Mann and Thompson, 1988; Polanyi, 1988; Grosz et al, 1995), much work has been done on automatically detecting

relationships between sentences and other discourse units (Marcu and Echibabi, 2002; Burstein et al. 2003; Chan, 2004). The discourse analysis was also extended to cross-document relationship modeling and exploration (Radev, 2000; Zhang et al, 2003).

However, the natural language processing approaches do not easily capture the underlying high level semantics that make topics, nor the relationships between them. To explore such relationships, people often attempt softer approaches: using analysis tools on the data and allowing people to evaluate the existence and types of relationships. Visualization is the most widely used techniques in such analysis.

Visualizing Relationships in Categorical Data

Visualization is used on large textual data sets primarily to show overviews or patterns in the data (Wise et al, 1995; Chen et al, 1998; Eick, 1994). However, with the exception of showing similarities, there has not been much work on revealing relationships between categories of textual data. Fortunately, many techniques of displaying relationships in non-textual data can be easily adopted for textual data. Below, we list several of these visualization techniques. One of them - Brushing & Linking - will be used in our study.

The Brushing & Linking technique (Becker and Cleveland, 1987; Ward and Martin, 1995; Stolte et al., 2002) is widely used in interactively indicating which parts of one data display correspond to that of another. When a system shows different views of the same set of data, the user can select and highlight a section of data in one view, and the system visually displays the distribution of the same set of data in another view. It has been used for categorical comparison to help people perceive differences between classification schemas (Graham and Kennedy, 2001; Munzner et al., 2003). The Mosaic display (Hartigan & Kleiner, 1981; Friendly 1994) visualizes n-way contingency tables. By portraying table entries as “tiles” whose areas are proportional to the value in the cell, it reveals pair-wise relationships between categorical variables (Friendly, 1999). The Parallel Set (Bendix, 2005) technique shows relationships among multiple categorical variables. Each variable is represented as a set of boxes representing its different categorical values, Each box is scaled according to the size of the corresponding category. The sets of categories for different variables are displayed side by side, with links among them showing the relationships between the variables.

System and Tools

For this study, the **Cosen**
system [Qu, 1995] was extended to support relationship exploration. Features supporting representation and manipulation of multi-faceted, nonexclusive categories were implemented, including: a) A special type of “link file”, which are pointers to other

documents. Using link files, we can assign a document into different categories without creating separate copies. b) Two distinct view options: a single-tree view and a double-tree view of the multi-faceted categories. In the double tree view, the two windows are laid side by side (Figure 1(a)), sharing the same underlying data structure, but having separate controls. Users can choose to view, expand, or collapse different categories, and conduct different searches in separate windows.

Moreover, special functionalities were designed respectively for the Overlap approach and the Search approach of relationship exploration.

1. 1. Brushing & Linking (B&L) functionality for the Overlap approach. We chose the Brushing & Linking technique to visualize the overlaps between categories. Compared with other visualization techniques, B&L allows user to set the visualization interactively, and can be directly implemented on the tree representations we have in the Cosen system. Using B&L, a user can choose a set of documents in one or more categories (Brushing) and see how they are distributed in other categories (Linking). The system highlights the originally selected documents and their appearance in other categories, thus displaying the overlaps between the original and other categories. The B&L functionality is available in both the single-tree and double-tree views. Figure 1(a) shows B&L in the double-tree view. A user selected file or folders in the left-hand-side tree window, and can examine the appearance of the corresponding documents under other folders in the right-hand-side tree window.
2. 2. Searching & Highlighting (S&H) functionality for the Search approach. To search over a category, the user should first select the category, type a query into the search field, and click the search button (Searching). If multiple keywords are entered, the “AND” operation is presumed among them. All documents containing the keywords will be highlighted in the categorical view (Highlighting) allowing the user to see the overlap between the search topic and the categorical topics. The Searching & Highlighting functionality is available in both the single-tree (Figure 1(b)) and double-tree views.

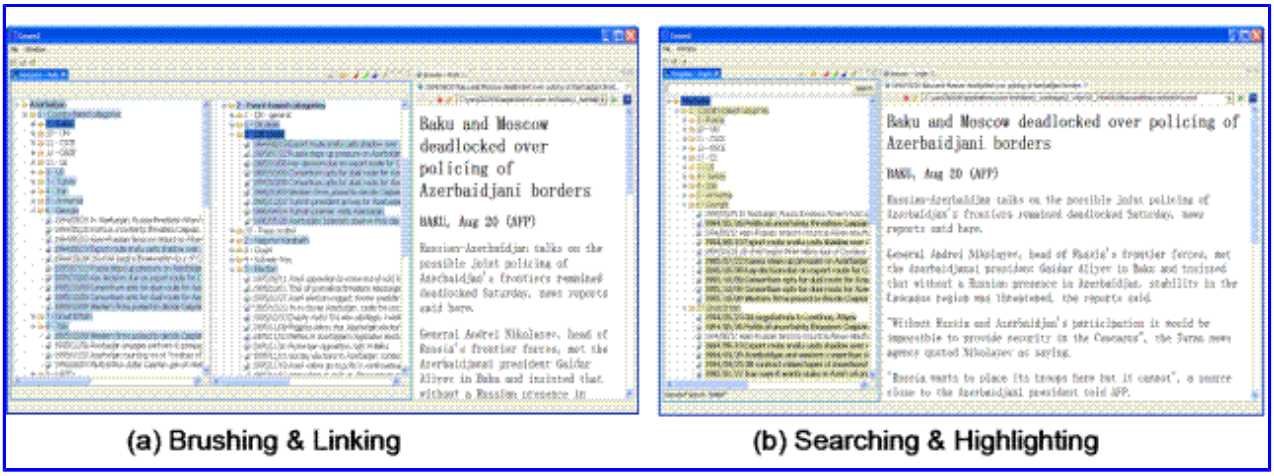

Figure 1. Functionalities supporting relationship exploration

Experiment Design

An experiment was designed to assess the effectiveness of the Overlap and Search approaches in relationship exploration. We developed a task requiring comprehending relationships in a set of news articles. The news articles were organized into multi-faceted non-exclusive categories such that the both approaches could be used on them. Questions about relationships between topics were asked. The score of the answer and the time spent on the task were used as measurements of users' performance.

Participants and Experiment Groups

Thirty nine graduate students from the University of Michigan, School of Information were randomly assigned to three groups. The Overlap group explored relationships by checking the overlaps between categories using B&L. The Search group explored relationships by searching on categories using S&H. The Control group used only the basic tree browser and the text reader functionality

During pilot testing, users preferred the double-tree view when using the B&L functionality, which allowed them to view the two intersecting categories side by side. However, they preferred the single-tree view when using the S&H functionality because they did not need to compare categories. We used these corresponding preferred modes in the experiment. The control group used the simple single-tree view.

Data Set

The data set was one hundred news articles on Azerbaijan taken from the Giga-Word corpus collected by the Linguistic Data Consortium. These articles were manually placed into two distinct categorical schemas. One was event oriented with categories such as "parliamentary election", "subway accident", etc.; the other was country oriented with categories such as "Russia", "Turkey", etc. The events and the countries were

selected by human judges who read all the news articles. The criterion of categorization was that, if an article mentioned a country or an event, it was put under that category. The categories within schema were not mutually exclusive: if an article mentioned several countries or events, it was placed under all of the relevant categories.. We asked participants to assume for the purposes of the study that the categorization was correct and complete, i.e., that everything in the data set relating to a country or an event was under the corresponding folder. The effects of this strong assumption are explored in the Discussion section.

Task and Performance Assessment

Participants were asked a sequence of questions about relationships between topics (countries and events) in the news set and used their assigned version of the system to help answer the questions. A time limit was set for each question. Time spent on each question was recorded and answers were graded by two scorers, with the mean value used in analyses.

Question Set

Below we show all the questions used on in the experiment. All questions were about relationships between topics that have corresponding categories in the data set, except the first two baseline questions (Q1 asked about a simple fact, Q2 asked about a relationship evolving topics that do not have corresponding categories). The Overlap group and the Search group may have advantages in answering these questions except the baseline ones. Inside the parentheses following the question ID (Q1-Q11), we have names of the questioned topics (mostly the same as the name of the associated category), and question types. All the italic words are names of existing categories. The question types will be explained after showing the list of questions.

Q1(Baseline question): Who was the president of Azerbaijan in 1994?

Q2(Baseline question): Azerbaijan had war on the territory of *Nagorno Karabakh* with which country?

Q3(*Russia-Coup*; Dot question; YesOverlap-YesAnswer question): Do the articles mention suspicions that *Russia* got involved in the *coup* in Azerbaijan in Oct, 1994?

Q4(*Georgia-Peace Process in Nagorno Karabakh*; Dot question; YesOverlap-NoAnswer question): Do the articles mention that Georgia got involved in the peace process in Nagorno Karabakh?

Q5(Armenia-Subway fire; Dot question; NoOverlap-NoAnswer question): Do the articles mention suspicions that *Armenia* got involved in the *subway fire* in Baku,

1995?

Q6(*Coups-Oil routes*; Dot question; NoOverlap-NoAnswer question): Do the articles mention that coups in Azerbaijan influenced the negotiation of *oil routes*?

Q7(Subway fire-Election; Dot question; YesOverlap-NoAnswer question): Do the articles mention any evidence that the subway fire was related to the parliament election two weeks later?

Q8(*Georgia-Turkey-Oil route*; Dot questions; YesOverlap-YesAnswer question; three-topic relationship question): In the choices of *oil route*, did *Georgia* and *Turkey* have shared interests or conflicting interests, or none of them?

Q9(*Election*-countries; Line question; including YesOverlap-NoAnswer, NoOverlap-NoAnswer sub questions): Do the articles mention any evidence that any of the following countries tried to influence the *election* in *Azerbaijan: Russia, United State, Turkey, Iran, Armenia*?

Q10(*Great Britain*-events; Line question; including YesOverlap-YesAnswer, YesOverlap-NoAnswer, NoOverlap-NoAnswer sub questions): According to the articles, *Great Britain* got involved in which of the following Azerbaijan issues: *Oil deal, Peace process of Nagorno Karabakh, Election, Subway fire, Coup*?

Q11(*United State-Russia*-events; Area question; including YesOverlap-YesAnswer, YesOverlap-NoAnswer, NoOverlap-NoAnswer sub questions; three-topic relationship question): In which of the following events did both *Russia* and *United State* get involved in: *Oil deal, Peace process of Nagorno Karabakh, Election, Subway fire, Coup*?

The question set contains various types of questions with different properties. Using them, we explore the strengths and weaknesses of each relationship exploration approach in answering different types of questions.

First, the questions may differ in complexity level. We illustrate this in Figure 2, placing existing topics (associated with categories) on both the x and y axes. There are questions asking about the relationship between two individual topics, whose answer resides in a “dot” (the crossover of two individual topics) in the diagram (“Dot” questions, Q3-8). There are questions asking about relationships between one topic and many other topics, whose answer resides in many “dots” over a “line” (“Line” questions, Q9, Q10). There are also questions asking about relationship between many topics and many other topics, whose answer resides in many “dots” over an “area” (“Area” question, Q11). Generally, “Dot” questions are easier than “Line” questions, and “Line” questions are easier than “Area” questions.

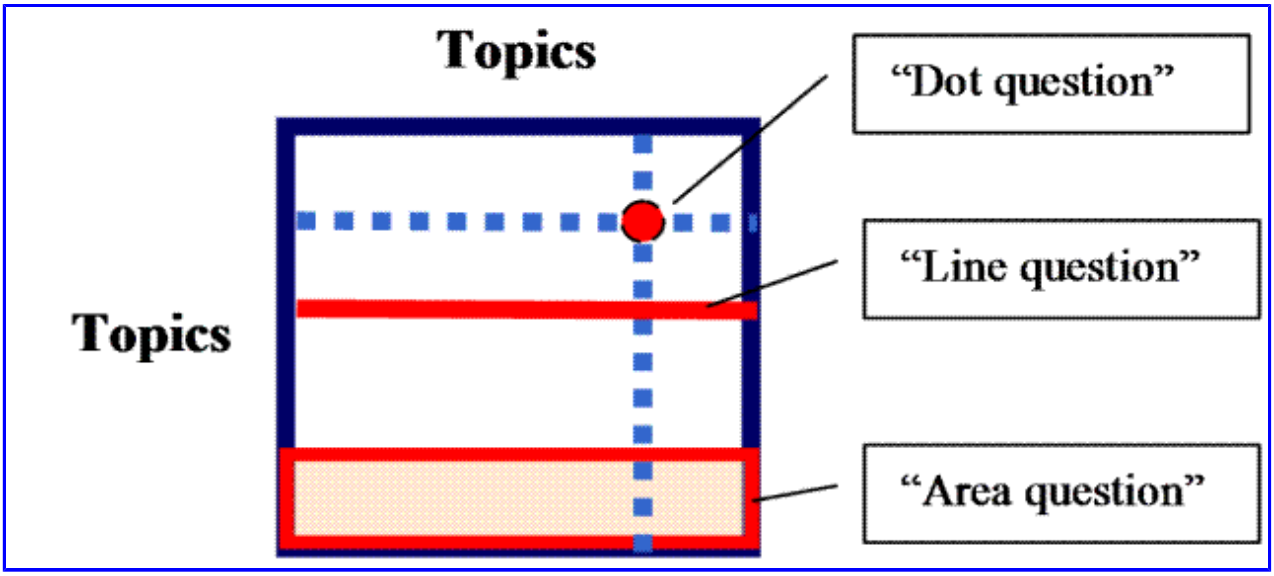

Figure 2. Different complexity of questions

We expected that both the Overlap and Search approaches would prove useful in answering "Dot" and "Line" questions. Notice that for the "Line" questions, both approaches can easily reveal the relationship between one topic (represented by a category or a query) and many other topics (represented by many categories). However, we were not optimistic about the "Area" questions because there is no easy way for either approach to compare different lines in the area.

In addition to varying complexity, the questions also differ in regards to the existence (or non-existence) of the overlap between relevant categories and the existence (or non-existence) of the questioned relationships. For some questions, the overlap of the relevant categories is not empty and the questioned relationship indeed exists, so that the answer to the question is "Yes" ("YesOverlap-YesAnswer" questions, Q3, Q8). For some questions, the overlap of the questioned categories is not empty, but the questioned relationship does not exist, so that the answer is "No" ("YesOverlap-NoAnswer" questions, Q4, Q7). For some questions, the overlap of the two relevant categories is empty. Based on our assumptions that everything related to a topic is put into the corresponding category, the questioned relationship does not exist. Therefore the answer is "No" ("NoOverlap-NoAnswer" questions, Q5, Q6). The "Line" or "Area" questions can usually be decomposed into a set of "YesOverlap-YesAnswer", "YesOverlap-NoAnswer", and "NoOverlap-NoAnswer" questions.

We expected both the Overlap and Search groups to have better performance than the Control group on "NoOverlap-NoAnswer" questions since ideally, people using the Overlap and Search approaches will see there is no intersection between the questioned topics without reading the articles. The performance of the Overlap and Search groups on "YesOverlap-YesAnswer" and "YesOverlap-NoAnswer" questions

may depend on the size of the overlap or search results compared with the size of the categories.

In the experiment, we also included questions about relationships involving more than two topics (three-topic relationship questions) to explore whether the approaches help with such questions.

Minimizing the Carry-Over Effect

The questions are not independent of each other. For example, both Q3 and Q6 ask about coups, both Q8 and Q4 ask about Georgia, and the "Line" and "Area" questions refer to many topics mentioned in the "Dot" questions. Therefore, answering one question may benefit answering another. We put the questions into four groups according to the question complexity: baseline questions (Q1-2), "Dot" questions (Q3-8), "Line" questions (Q9, Q10), and "Area" Questions (Q11). To minimize the carry-over effect, we asked the questions in order of increasing complexity (the greater the complexity of a question the more topics it covers, thus it has a bigger carry-over effect than other questions). Within each group, the questions were asked in a random order.

Data Collection

In the experiment, we not only asked participants to give simple "yes" or "no" answers, but required them to provide evidence to support the answer. We recorded the answer, the evidence, as well as the elapsed time for each question. Additionally, a logging tool within the system recorded various activities conducted by the participant while answering each question, including: reading articles, using B&L or S&H operations, etc.. In a short interview after the experiment, we asked participants about difficulties in using the system and in finishing the task.

Data Analysis and Results

In this section, we first assess the performances of different groups using the time they spent and the scores they have on the questions. Then, we perform a closer investigation into what different groups did to answer the questions, in order to explain their performances in the experiment. Several interesting issues are raised as well.

Performance Assessment

Performance assessment based on time

Figure 3 shows the average time spent on each question by each group. Overall the Overlap group spent significantly less time than the Search group and the Control group ($p=0.04$, $p=0.008$ respectively, one-way ANOVA, between groups, multiple comparison). There is no significant difference between the Search group and Control group.

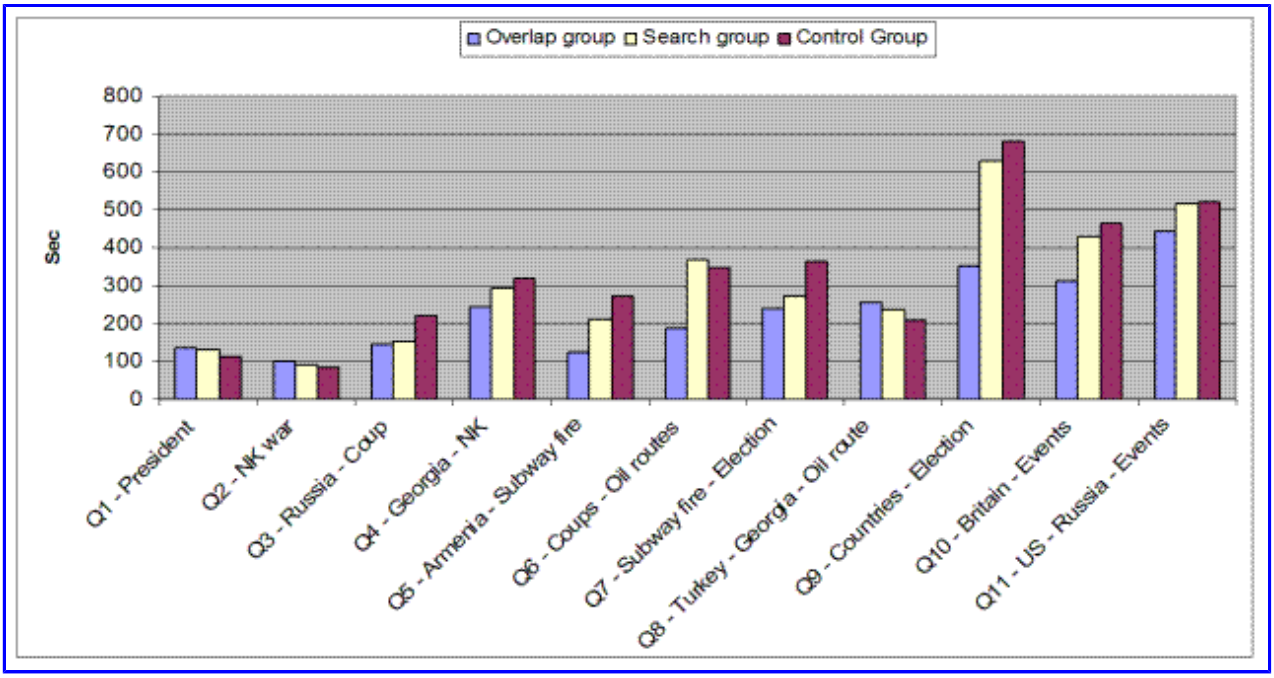

Figure 3. Average time spent on each questions by each group

A further examination on the data showed the performance of the three groups varied in answering different types of relationship questions.

First, checking participants' performances on question groups of different difficulty levels, we found that the Overlap group outperformed the Control Group on "Dot" questions(Q3-8) ($p = 0.02$, one-way ANOVA) and "Line" questions(Q9-10) ($p<0.001$, one-way ANOVA) as we predicted. However, the Search group was not significantly faster than the Control group on either "Dot" questions or "Line" questions. There is no significant difference among groups on the baseline questions (Q1-2) and "Area" questions(Q12).

Second, in the set of "Dot" questions, we analyzed the group performances on "YesOverlap-YesAnswer", "YesOverlap-NoAnswer" and "NoOverlap-NoAnswer" questions respectively. For "NoOverlap-NoAnswer" questions (Q5, Q6), the Overlap group spent significantly less time than the Search group ($p=0.03$) and the Control group ($p=0.01$) respectively, while there was no significant difference between the Search and Control groups. For "YesOverlap-NoAnswer" questions (Q4, Q7), the Overlap group showed the trend of spending less time than the Control group ($p=0.1$), while there was no significant difference between the Search group and the Control group. However, for Q7, we found two participants in the Control group spent out the limited time for the question (10 minutes). Were there no limit on the time, there might have been a significant difference between the Overlap and Control groups. For "YesOverlap-YesAnswer" questions (Q3, Q8), there was no significant difference among groups.

Third, we checked questions asking about relationships involving three topics (Q8, Q11) and there was no significant difference between groups.

In our experiment, a time limit was set for each question (5 minutes for Q1-2; 10 minutes for Q3-8; 20 minutes for Q9-11). However, the caps are high enough so that few people actually used up their time in the experiment. Were the limit removed, our findings would remain the same except for Q7, as discussed previously.

Performance assessment based on scores

In our experiment, the score of the answers was another measure of the performances. We graded participants' answers for each question. In an attempt to reflect the relative difficulty of the different question types, the baseline questions were worth 3 points apiece; the "Dot" questions 5 points apiece; and the "Line" and "Area" questions 10 points apiece. The average scores of each group for each question are shown in Figure 4.

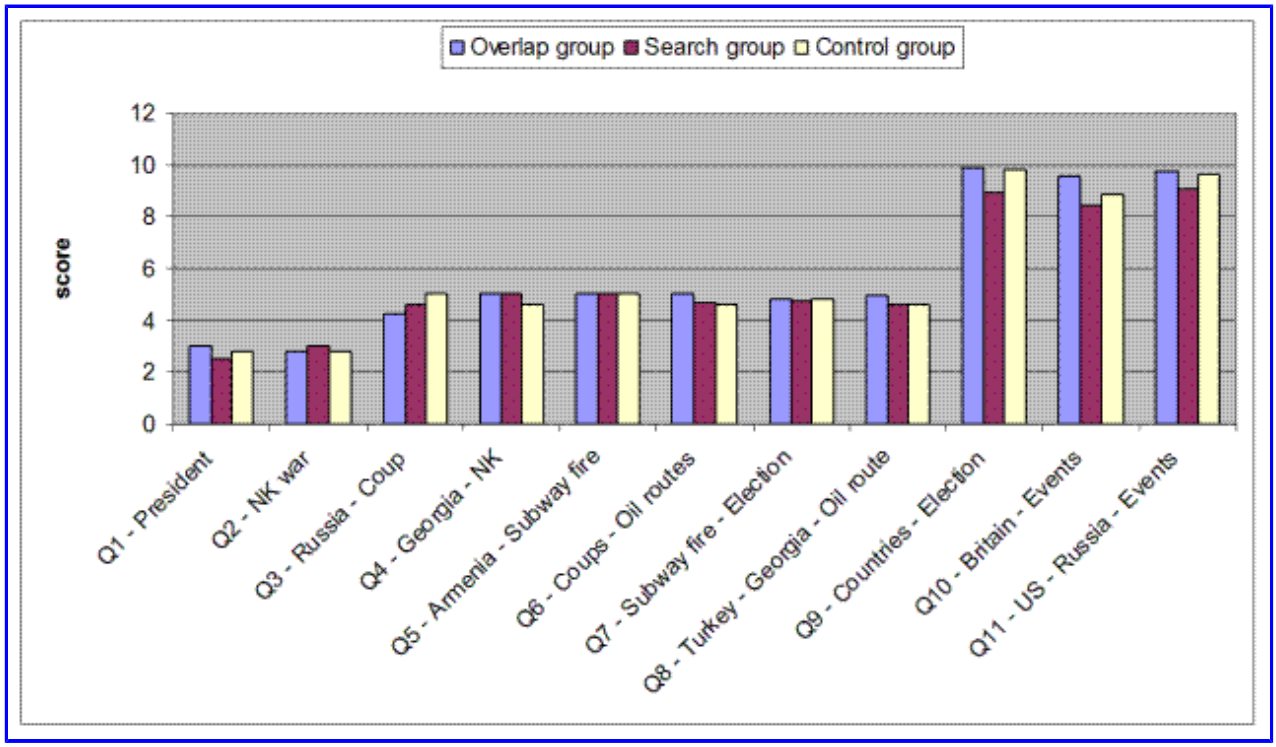

Figure 4. Score comparison of the three groups on each question

Since we set high caps on time limits, participants had fairly high scores. The differences among group scores are not as obvious as the time differences. Overall, the Overlap group had higher scores than the Search group ($p=0.05$, one-way ANOVA). However there was no significant difference between the Overlap group and Control group, or the Search group and Control group.

After closely inspecting each question type, the only type of questions which induced

significant differences between the groups were the “Line” questions (Q9, Q10). To our surprise, both the Overlap and Control groups had significantly higher scores than the Search group (with $p<0.001$, $p=0.03$ respectively).

In summary, the Overlap approach did help in relationship exploration by shortening the time to answer “Dot” and “Line” questions. It was most efficient in answering “NoOverlap-NoAnswer” questions. Conversely, the Search group did not improve the performance in terms of either time or score.

Next, by taking a close look at what participants did to answer the questions, we explain why the Search approach did not help in this task and raise several interesting issues in relationship exploration.

Close Investigation of the Search Group

In this section, we seek explanation of the inefficiency of the Search approach in our experiment by analyzing the activities of participants in relationship exploration.

Figure 5 shows the number of articles read by each group for each question in the experiment. We can see that the Search group read fewer articles than the Control group, as we expected (because the size of search results is smaller than the original category). However, comparing Figure 5 with Figure 3, we see that the differences between the Search group and the Control group on the time spent on each question are much smaller than the differences on the number of articles read. So, something other than reading the articles may have slowed the Search group.

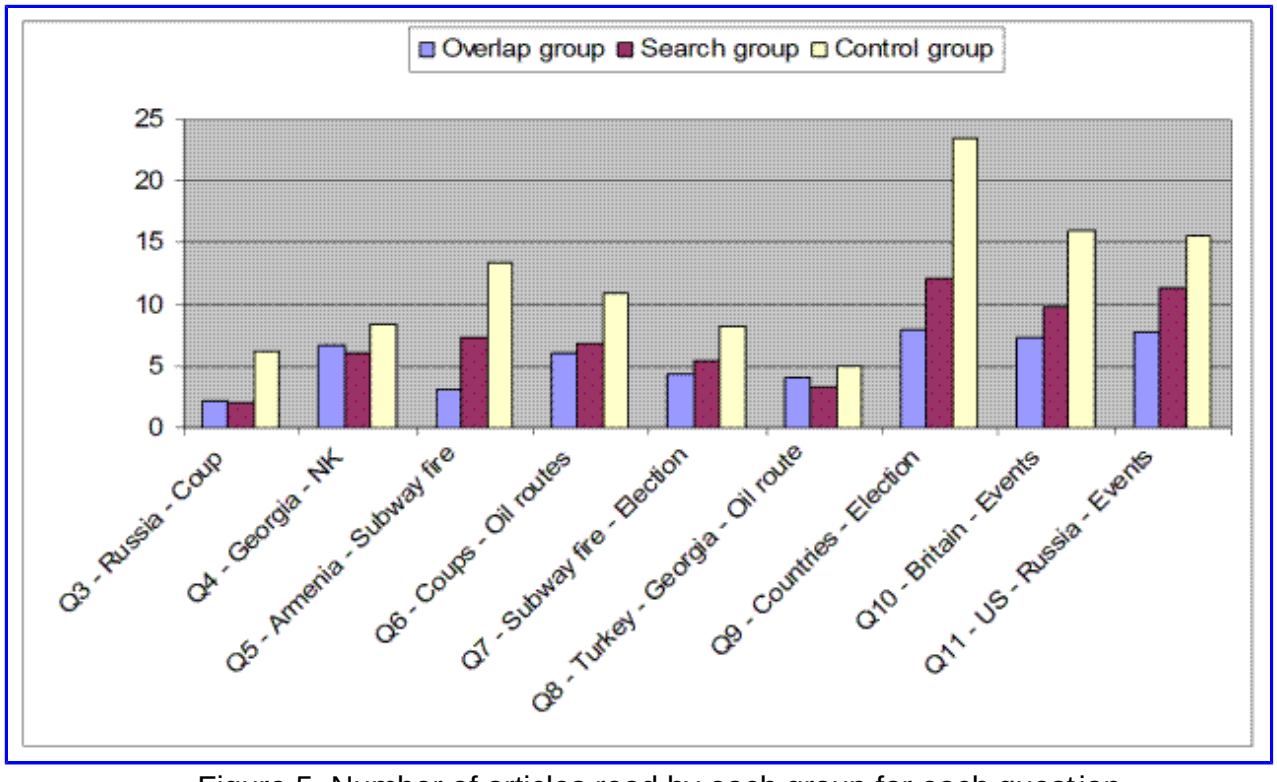

Figure 5. Number of articles read by each group for each question

One reasonable explanation is that the Search group spent extra time proposing queries and skimming through the article titles appearing in the search results. Figure 6 shows the average number of queries (with duplicates) submitted on each question by the Search group. For those questions with a sizable difference between the number of articles read by the Search and Control groups and with a much smaller difference between the time spent by the two groups (Q5, Q6, Q9, Q10 and Q11), the number of queries submitted by the Search group was large. Therefore, participants in the Search group probably spent considerable amounts of time searching, thus losing the time gained by reading fewer articles.

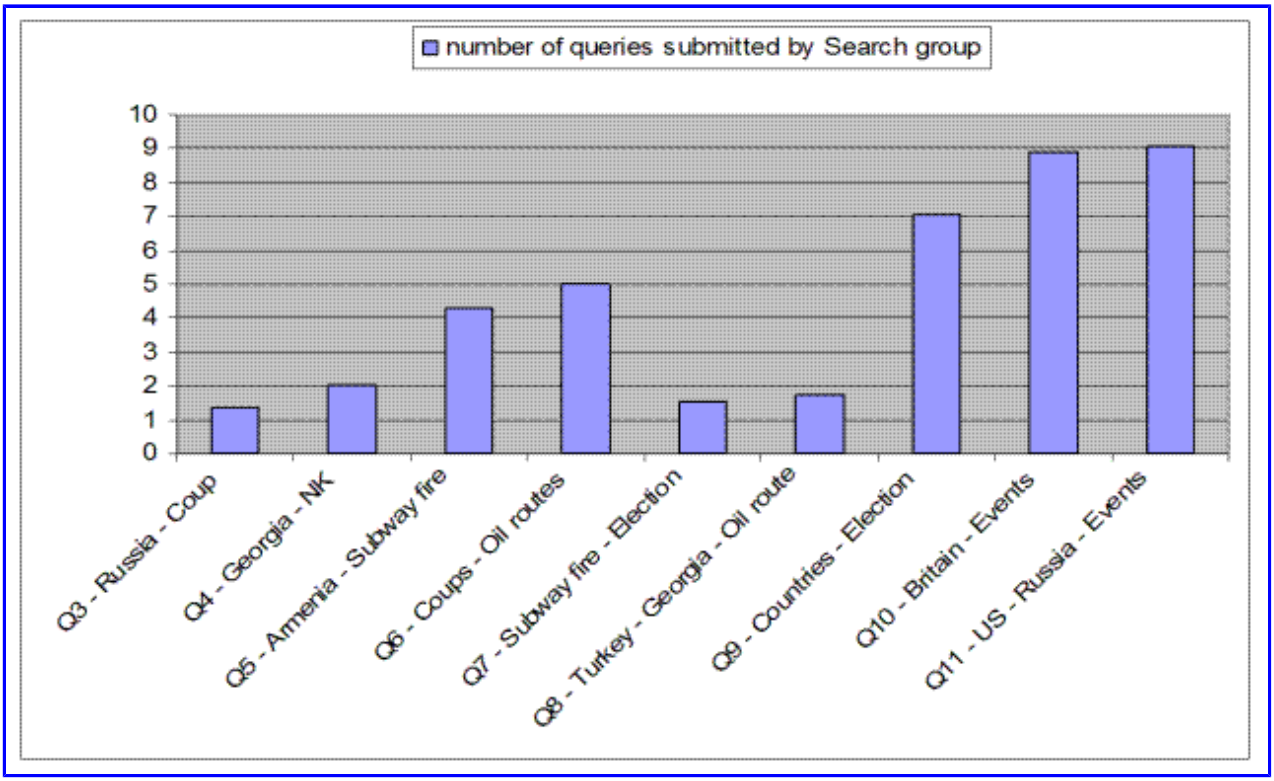

Figure 6. Number of queries asked on each questions

A closer examination of the log data indicates that participants in the Search group faced the Vocabulary Problem (Furnas et al., 1987): since there may be many different words used in the news articles referring to the same topic, one keyword search on that topic may not retrieve all the relevant document. Additionally, it is usually difficult for a person to recall all the synonyms on a topic and even harder for him to know whether he has exhausted all possible keywords. In the experiment, we observed that some participants in the Search group kept trying different search keywords, hesitating to stop searching. At the same time some participants stopped their search too early such that they did not try the “correct” keyword, and thus missed the evidence. This is probably

the reason for the lower score of the Search group on the “Line” questions (Q9, Q10). For example, one subject searched “Great Britain” for Q10 (Britain vs. several events) and missed the article containing evidence for one of the choice, which used only “Britain” but not “Great Britain”.

Close Investigation of the Overlap Group

The Overlap group outperformed the Control group as we predicted. In this section, a detailed analysis of the reading load for the Overlap group and the Control group raise several interesting issues behind the scenes.

Table 1. The estimated number of articles to be read and the real number read to answer each question

	Overlap group - estimated	**Overlap group - real**	**Control group - estimated**	**Control group - real**
Q3 *Russia-Coup*; Dot; YesOverlap-YesAnswer	<9	2.2	<17	6.3
Q4 *Georgia-Peace Process of Nagorno Karabakh*; Dot; YesOverlap-NoAnswer	4	6.8	10	8.3
Q5 *Armenia-Subway fire*; Dot; NoOverlap-NoAnswer	0	3.2	17	13.3
Q6 *Coups-Oil routes*; Dot; NoOverlap-NoAnswer	0	6	9	10.8
Q7 *Subway fire-Election*; Dot; YesOverlap-NoAnswer	2	4.3	17	8.2
Q8 *Georgia-Turkey-Oil route*; Dot; YesOverlap-YesAnswer	<9	4.1	<9	4.9
Q9 *Election*-countries; Line	2	7.9	17	23.4
Q10 Great Britain-events; Line	<13	7.3	14	16
Q11 *United State-Russia*-events; Area	<34	7.8	<43	15.6

Table 1 shows the estimated number of articles that must be read to answer each relationship question, and the number of articles actually read by the Overlap and the Control group. For the Overlap group, the number of articles that must be read depended mainly on the number of overlapping articles between different categories. For “YesOverlap-YesAnswer” questions (Q3, Q8), the number of articles that must be read is less than the number of the overlapping articles because the participant may

have found the articles containing the evidence before reading all overlapping articles. For “YesOverlap-NoAnswer” (Q4, Q7) questions, the number of articles that must be read is equal to the number of overlapping articles, because participants needed to check all the overlapping articles to make sure no evidence existed. For “NoOverlap-NoAnswer” (Q5, Q6) questions, since there was no overlap between the two relevant categories, the participant could answer “No” to the question without reading any articles. For more complicated questions (Q9-11) we decomposed the questions into “YesOverlap-YesAnswer”, “YesOverlap-NoAnswer”, “NoOverlap-NoAnswer” sub-questions and estimated the reading work load. For the Control Group, the work load is estimated by counting the number of articles in the relevant category of the smallest size. For example, for Q3 (Russia-Coup), the “Russia” category contains 38 articles, the “Coups” category contains 17. The answer of the question resides in both categories (in the overlap), therefore, the participant needs only to go through the category of smaller size to answer the question. And since the evidence exists for Q3, the estimated work load for the Control Group is less than 17.

One interesting observation from Table 1 is that participants in the Overlap group read more than necessary for many questions (Q4, Q5, Q6, Q7, and Q9). Furthermore, those questions had one important commonality: they all had the “No” answer (Q9 has sub-questions that have “No” answers). It appears when there is no overlap between relevant categories or when participants could not find the questioned relationship in the overlap, they were hesitant to say no and read more articles outside the overlap to check for the existence of the relationship. Our interview with the participants confirmed this conjecture. Moreover, we found that this unwillingness to stop resulted from the lack of trust of the categories built by other people (the experimenters, in this case) or a system. When no relationship was found, people often suspected some important information was not put in the correct category, and thus was not shown in the overlap. The lack of trust in the categorizations may also affect the performance of the other two groups. For example, in Table 1, we see participants in the Control group also sometimes read more than necessary (Q6, Q9, Q10).

Another interesting observation is that, for the Control group, there are several questions in which people read less then necessary (Q4, Q5, Q7). We suspect this resulted from the carry-over effect - some articles had been read when answering previous questions so that they did not need to be read again. This carry-over effect may be more obvious in the Control Group than in other groups because those participants read more for each question. For many questions, they theoretically needed to read through the whole category to answer the question. Therefore, the fact that they went though a category for a previous question could very possibly benefit later questions related to the same categories. We also suspect that the carry-over effect not only decreased the number of articles read in latter questions, but also shortened the average time to read each article

as time passed by. Further investigation is needed to understand the carry-over effect.

Discussion

In the experiment, the Overlap approach helped people answer relationship questions faster, while the Search approach did not show the same effectiveness. In this section, we take a closer look at the underlying differences between the two approaches.

First, how quickly and completely the two approaches can reveal topic collocations are influenced by different factors. For the Overlap approach, the quality of the categories is a crucial factor. If the categories are well built (each matches a topic, containing everything relevant and nothing irrelevant) one Brushing & Linking operation (Overlap approach) would show all the collocations of the questioned topics. If the categories are not well built, the effectiveness of this approach will be impaired. For the Search approach, besides the quality of the categories, the vocabulary problem limits its effectiveness heavily. When there are many synonyms for a topic, it takes several Searching & Highlighting operations on a category to reveal all the topic collocations. Failure to exhaust all possible keywords could result in missing topic collocations.

Second, the different characteristics of the two approaches mentioned above lead to different human behaviors in using the approaches. In our experiment, we found that relationship exploration is much more than simply finding the correct categories or keywords and performing the B&L or S&H operations. It is a continuous process involving many decisions by the user. The two most important decisions that affect people's performance are: whether to trust the result when no relationship is found, and whether to continue the exploration.

1. Trust or distrust the result. When the system shows the collocations of topics, the user checks the topic collocations for relationships. If he finds the questioned relationship, the task is finished. If not, he needs to decide whether he trusts the results or not - is there really no such relationship? Or did the system miss some important information due to the quality of the categories or the Vocabulary Problem? This decision is made based on people's assessment of the category quality, and estimation of the coverage of the keywords. If they believe the category is not well built, or there are many synonyms for the topics, then they may suspect the system failed to show the relationships as a consequence.
2. Continue or stop the exploration. When a person suspects the relationship exists but it is not shown by the result of B&L or S&H operations, he needs to decide if he wants to continue the exploration and what to do next. What people will do and how much more time and effort they will spend depends on the costs/benefits of their choices. For the Overlap approach, the relationship information can be missed when one of the relevant categories does not include all the relevant

articles. To find the missed information, people could check the non-overlapping part of the relevant categories (with relatively low cost), and then check other places (with relatively high cost). For the Search approach, relationship information can be missed either because the previous queries have not covered all the topic collocations, or because the category dose not include all the relevant articles. Therefore the user could choose to try new queries on the relevant category (with relatively low cost) or check documents outside the relevant category (with relatively high cost).

In our experiment, when participants could not confirm the questioned relationship, they intended to continue the exploration. We observed people in the Search group conducted more searches, and people in the Overlap group read more articles. Participants in the Search group were more hesitant to stop since they were not sure whether they had exhausted the synonyms, and the cost of conducting new searches was relatively low. Although they spent longer time in the task compared to the Overlap group, some people still missed the correct answer because they stopped trying different queries too early.

In summary, the characteristics of the two approaches lead to different human behaviors in using the approaches. Together, they account for the dissimilar performances in our experiment.

Design Implications

The understanding of the relationship exploration processes leads to several design implications:

First, the quality of the categories and the Vocabulary Problem were the factors heavily influencing the effectiveness of the Overlap and Search approaches. A system implementing these approaches for relationship exploration should provide information about these factors so that users can correctly assess the quality of the result. For example, a system could inform users how the categories are generated, informing users as to the quality of the categories. The system should also remind users about the existence of the Vocabulary Problem when they search, or even provide possible synonyms.

Second, the experiment shows that relationship exploration does not end after the first try. People may not be satisfied with the current results and decide to continue. Thus, a relationship exploration system should provide support for this continuous process. For example, it could record intermediate results and activity histories (e.g., a list of read documents or submitted queries) to help users track their progress.

Third, the relationship exploration systems are often aimed at helping users quickly

answer questions without significant reading. However, more background reading will give people more contextual knowledge and better comprehension of the data, which may bring about long-term benefits. Therefore, it will be helpful to add tools to extract contextual information and summarize background information.

Conclusion and Future Work

Perceiving relationships is an important step leading toward a deeper understanding of textual data. In this work, we explored two approaches of using category information in relationship exploration of textual data: the Overlap approach and the Search approach. We conducted an experiment to investigate their effectiveness in solving different types of relational questions. Our results show that the Overlap approach indeed helped in exploring relationships in categorical data. At the same time, the Search approach did not improve users' performance mainly because of the Vocabulary Problem.

This work is a first step in investigating how to design tools to facilitate exploring relationships in textual data using category information. There remain several topics for future exploration, including: 1) Exploring more efficient approaches to help with answering complicated relationship questions, such as "Area" questions and three-topic relationship questions; and 2) Testing the effectiveness of the Search approach on a large scale dataset, to see if the time saved by reducing the reading load will exceed the time spent on conducting searches.

ACKNOWLEDGMENTS

This work was supported by a grant from the National Science Foundation (IIS-0325347-ITR). The authors would like to thank Soo Young Rieh, Lingling Zhang, Jun Zhang, and Nikhil Sharma for their various thoughts and help.

References

Becker, R. A., & Cleveland, W. S. (1987) Brushing Scatterplots *Technometrics* 29(2):127--142

Bendix, F., Kosara, R., & Hauser, H. (2005) Parallel Sets: Visual Analysis of Categorical Data *Proceedings of IEEE InfoVis '05* pp. 133-140

Burstein, J., Marcu, D., & Knight, K. (2003) Finding the WRITE Stuff: Automatic Identification of Discourse Structure in Student Essays *IEEE Intelligent Systems* 8(1): 32-39

Chan, S. W. K. (2004)
Automatic discourse structure detection using shallow textual continuity *International Journal of Human-Computer Studies* 61(1): 138-164

Chen, H., Houston, A. L., Sewell, R. R., & Schatz, B. R., (1998) Internet Browsing and Searching: User Evaluations of Category Map and Concept Space Techniques *Journal of the American Society for Information Science* 49(7): 582-603

Eick, S. G. (1994) Graphically displaying text *Journal of Computational and Graphical Statistics* 3: 127-142

Friendly, M. (1994) Mosaic displays for multi-way contingency tables *Journal of the American Statistical Association* 89: 190-200

Friendly, M. (1999) Visualizing Categorical Data In Sirken, Monroe G. et al. (Eds.) *Cognition and Survey Research* 319-348. New York: John Wiley & Sons

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

Graham, M., & Kennedy, J. (2001) Combining linking & focusing techniques for a multiple hierarchy visualisation *Proc. of IV 2001 - 5th International Conference on Information Visualization* pp 425-432

Grosz, B. J., Joshi A. K., & Weinstein, S. (1995) Centering: a framework for modeling the local coherence of discourse *Computational Linguistics* 21(2):203--255

Hartigan, J. A., & Kleiner, B. (1981) Mosaics for contingency tables In W. F. Eddy (Ed.) *Computer science and statistics: Proceedings of the 13th symposium on the interface* pp. 286-273. New York: Springer-Verlag

Hearst, M. A. (1998) Automated discovery of WordNet relations In Christiane Fellbaum (Ed.) *WordNet: An Electronic Lexical Database* MIT Press

Hindle, D. (1990) Noun classification from predicate-argument structures *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics* pp 268-275

Mann, W. C., & Thompson, S. A. (1988) Rhetorical structure theory: toward a functional theory of text organization *Text* 8(3): 243-281

Marcu, D., & Echihabi, A. (2002) An unsupervised approach to recognizing discourse relations *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)* pp 368-375

Munzner, T., Guimbretiere, F., Tasiran, S., Zhang, L., & Zhou, Y. (2003) TreeJuxtaposer: Scalable Tree Comparison using Focus+Context with Guaranteed Visibility *ACM Transactions on Graphics* 22(3): 453-462

Polanyi, L. (1988) A formal model of the structure of discourse *Journal of Pragmatics* 12: 601-638

Qu, Y. (2003) Sensemaking-Supporting Information Gathering System *Extended Abstract of Conference on Human Factors in Computing Systems (CHI 2003)* pp906-907

Radev, D. (2000) A common theory of information fusion from multiple text sources, step one: Cross-document structure *Proceedings of the 1st ACL SIGDIAL Workshop on Discourse and Dialogue, Hong Kong, October 2000*

Stolte, C., Tang, D., & Hanrahan, P. (2002) Polaris: A System for Query, Analysis and Visualization of Multi-dimensional Relational Databases *IEEE Transactions on Visualization and Computer Graphics* 8(1): 52-65

Ward, M. O., & Martin, A. R. (1995) High Dimensional Brushing for Interactive Exploration of Multivariate Data *Proceedings of Visualization '95* pp. 271-278

Wise, J. A., Thomas, J. J., Pennock, K., Lantrip, D., Pottier, M., & Schur, A. (1995) Visualizing the non-visual: Spatial analysis and interaction with information from text documents *Proceedings of the Information Visualization Symposium* pp. 51-58

Zhang, Z., Otterbacher, J., & Radev, D. (2003) Learning cross-document structural relationships using boosting *Proceedings of the Twelfth International Conference on Information and Knowledge Management (CIKM 2003)* pp 124-130