

BEYOND WORD RELATIONS

SIGIR '97 Workshop

Organizer: Beth Hetzler, Pacific Northwest National Laboratory

Workshop participants: Carol Bean, NLM; Benjamin Ben-Ami, WISDOM; Archana Bhandari, Eastman Kodak; Byron Dom, IBM Research; Garrett Dworman, University of Penn.; Rebecca Green, University of Maryland; Tom Hewett, Drexel University; Noriko Kando, NACSIS, Japan/Lancaster University, UK; Susan Malveau, Oregon Health Sciences University; Curtis Means, Office of Research & Development; Sung Myaeng, Chungnam National University; Lucy Nowell, Virginia Tech; Jody Palmer, Open Text Corp.; Russ Rose, Office of Research & Development ; Hinrich Schuetze, Xerox PARC; Jim Thomas, Pacific Northwest National Laboratory ; Xiang Tong, Claritech Corp.; Charles Wayne, NSA

Abstract

Many information retrieval systems identify documents or provide a document visualization based on analysis of a particular relationship among documents — that of similar topical content. But there may be layers of other less apparent and less traditional relationships that are useful to the user. Exploring this other information was the subject of this workshop, with a focus on identifying new non-traditional relationships. An initial taxonomy was introduced and fleshed out during the workshop.

Introduction

When an information retrieval system answers a user query, the traditional approach is to find documents where the topic matches the topic of the query. However, users may care about many aspects of the documents other than matching topic. In a paper presented earlier in this conference, Fidel and Crandall showed that user criteria for document relevance included elements such as whether the document included hard data, was written on a technical vs. non-technical level, or confirmed what the user already knew [Fidel]. Candidates for potentially useful relationships among documents can be based on other aspects of the semantic contents, such as level of detail; on attributes of the documents, such as author or source; on whether the documents reference or quote one another; on whether documents have an influence on each other, such as a cause-and-effect relationship; and so on.

Visualization can provide a method to help users explore and understand such relationships among a collection of documents. A frequent visualization technique for relationship networks is a node and link diagram, commonly used in the field of link analysis [P1000]. However, large and complex node/link diagrams can lead to cluttered displays and other related problems which may add more confusion than clarity [Eick].

The goal of the workshop was to significantly enhance our understanding of the relationships and associations among documents by

- Identifying possible semantic relationships among documents
- Categorizing those relationships
- Identifying visualization and application possibilities

Presentations

The morning comprised four invited presentations selected to provoke thought and discussion.

Russ Rose, Office of Research and Development. Russ provided background on intelligence analysis tasks and challenges in sifting through large numbers of documents to select pieces of relevant information, where the pertinent relationships might be complex and hard to grasp. He also overviewed a

number of approaches currently being explored to visualize textual information. None of these visualizations currently addresses the potential richness of relationships that this workshop explored.

Tom Hewett, Drexel University. Based on the premise that understanding how the mind works should help us develop better systems, Tom's presentation involved interactive demonstrations to illustrate points about how people make and use associations and relationships in memory tasks. Using the session attendees as (usually) willing guinea pigs, he demonstrated that words associated with one another are more easily recalled than ones that are not. He showed how "errors" can illustrate various association methods the mind is using. He demonstrated that people may even remember associated items better than the target items they were asked to remember.

Rebecca Green, University of Maryland; Carol Bean, National Library of Medicine. Rebecca and Carol described their work, which examined the validity of the traditional assumption that the topic of a query should match the topic of those documents which are relevant to that query. In one study, they identified 33 topic-based relationship types, grouped into the following three larger relationship types:

- Topic-matching relationships (these accord with the traditional view)
- Hierarchical topic relationships — for example, where the topic of a relevant document is more specific than the topic of the query
- Structural topic relationships, where the query and relevant document topics are interrelated through one or more conceptual structures — for example, if the topic of the query is "debt," the topic of one relevant document might be "payment schedules," while the topic of another might be "compounding of interest"

Within their study, structural topic relationships were more common than hierarchical topic relationships, which were in turn more common than strict topic matching relationships.

Lucy Nowell, Virginia Tech. Lucy described the Envision system, which helps users visualize and explore attribute-based relationships among a set of documents resulting from a user query. For example, the user may select from a number of display options to portray documents' date of publication, author, relevance to query, etc. She summarized various sets of published guidelines regarding how to represent such information. The user studies she has performed with Envision have reinforced some of these guidelines and challenged others.

Identifying Relationships

The afternoon began with a brainstorming session to identify as many types of relationships as possible. To help facilitate the brainstorming, draft relationship categories were suggested in advance. However, participants were free to suggest new categories or relationships that did not fit in any of the suggested ones.

The group identified numerous relationship types. Although these types are most familiar in the context of text document-to-text document relationships, many are pertinent to other media, such as images or video, and some are more relevant to non-text entities. The categorization seemed to work fairly well as a starting point, although some relationships were seen to fall into more than one category, and some categories seemed to overlap.

Following the brainstorming, the participants broke into three groups, each of which explored a particular category. Below is a brief summary of each category, with examples of the relationships identified. The categories that were brainstormed only and not expanded by small group discussion are presented first.

Word-Based Relationships

These relationships are fairly traditional and familiar. Example relationships placed in this category include

- Same word co-occurs (disambiguation is important here)
- Same vocabulary is used in both documents
- Images similar between two complex documents

Attribute-Based Relationships

Many traditional relationships are included here, as well as some less commonly considered ones. Although listed in terms of matching attributes — such as “same author” — these could also be useful as close matches. The suggested relationships can be grouped into a number of sub-categories, such as

Attributes having to do with the creation of the document, e.g.

- Same author or source
- Same place of origin/nationality

Attributes of the content, e.g..

- Same level of detail
- Same style (or technique, as in photography or cinematography)

Attributes of the construction tools, e.g.,

- Same media (in music, could be same instruments)
- Same language

Categorizations outside of the document itself, e.g.,

- Same genre
- Included in the same collection

Document-To-Document Hierarchical Relationships

These are situations where one document (image, video, etc.) is a superset or subset of the other.

Examples are

- *A* is subpiece of *B*
- *A* is developmental form/stage of *B*

Document-To-Document Topological Relationships

This category is a conceptual extension of hierarchical relationships. Many kinds of relationships were placed here — perhaps indicating a need for further exploration and categorization. They could be grouped into the following sub-categories:

Conceptual equivalents, e.g.,

- *A* is translation of *B*
- *A* is transcription of *B* (audio)

Commentary or further work, e.g.,

- *A* indexes *B*
- *A* updates/corrects *B*
- *A* is a lecture on test while *B* is a discussion of test results

Sequential relationships, either physical or conceptual, e.g.,

- *A* follows *B* in a sequence
- Memo trail of a project/engagement (perhaps cross-departments)
- Right of, left of (e.g., photos of adjacent objects)

Relationships between the document sources

- Marketing department’s documents vs. Research department’s documents

Documents related to a common event

- Permission to travel to a conference vs. expense report filed afterward

Document-to-Document Influence Relationships

This category was meant to include situations in which one document affected the writing of another. There was an apparent overlap between relationships here and some of those placed in the previous category, indicating a need for further exploration.

Resulting

- Newsgroup question and answer
- Cause/effect
- A is one of the consequences of B

Positive response

- A builds on/expands work in B
- A draws conclusions from premises in B
- A substantiates findings of B
- A supports (lends support to assertions in) B

Negative response

- A contradicts B

The three categories that were chosen for further exploration after the brainstorming session were Topic-Based Relationships, Usage-Based Relationships, and “Other.” Below is a brief summary of these categories, example relationships identified for each, and additional notes based on the small-group discussions.

Topic-Based (or Meta-Topic-Based) Relationships

This category is a traditional basis for many systems. Some non-traditional extensions were identified, e.g.,

- A’s topic is subset of B’s topic
- Same moral

The small group focused on topic-based relationships that were not as apparent as those relationships that are commonly emphasized. They discussed situations in which document groupings (or clusters) are related, even though individual documents chosen from the two groups might not be related. They also discussed situations in which secondary topics of one document (or document set) might match secondary topics of a second document. For example, Don Swanson’s work has demonstrated several instances in which two bodies of literature that are related by subtopics but do not cite each other can actually lead to new discoveries in potential medical treatments [Swanson]. To help users identify and understand such less-apparent relationships, the group noted a need for a flexible interactive system that would allow users to expand or collapse the level of detail and to hide or display links as desired.

Usage-Based Relationships

These are relationships having to do with document users or with tasks related to the documents. Examples include

- All papers my boss has read
- Documents that were highly rated according to a given user’s profile
- Documents that share a common workflow

The discussion group that expanded these relationship ideas addressed user profiles and task workflow as useful applications and methods for expanding the ideas.

By “user profile,” we mean a method for specifying selection criteria for documents, so that it can be automatically decided which documents are likely to be useful to the user. It’s interesting to note that a profile might make use of relationships (“give me documents with the same source as this one”) and might also generate relationships (“all documents that score highly on Rebecca’s profile”).

For purposes of a user profile, the various relationships and criteria might be summarized in a score that shows the likelihood of a document being useful to the user. The user might want to modify a threshold value, determining which documents to show. Or a person might want a visualization that shows which

particular set of documents is most likely to be useful to a particular set of users. The group envisioned an illustration such as a terrain map with peaks where document/user matches were particularly high.

The group also discussed task workflow applications. Attributes such as criticality and time-dependence, inter-document relationships such as sequencing or prerequisites, and decisions about documents such as whether they are “done” or “not done” become important in this context.

Other

These were relationships that do not fit easily into a category. Examples include

- *A* links to *B*/*A* cites *B*
- Relationship suspected but not revealed
- Non-relationship/null-set/not related
- *A* is an enlargement of part of *B* (e.g., maps)
- *A* shows another perspective (as in pictures that show a different angle) of objects in *B*

The discussion group that expanded this category began by discovering a pattern in the set of ideas placed here. Users begin with a goal, such as discovering the reason for a (set of) generic relationships, or perhaps seeking to understand the absence of relationships in a particular subset of documents. The user creates a problem representation, such as the mathematical representation of graphs. Alternate representations can lead to alternate perspectives and new insights. The user also needs a set of tools to aid in creating and manipulating the representation. Useful tools might include sequencing capabilities (over time or other potential orders), variable granularity (the ability to view more or less detail, to summarize or expand), and animation. Visualization ideas discussed included

- Proximity or clustering
- Color/brightness
- Might use concept of overlap (as in Venn diagrams) or nesting
- Because both linkages and non-linkages are important, might allow capability to toggle between or to cross-fade between them

Conclusion

Both the variety and the number of relationships identified in this workshop were impressive. It's clear that a large number of relationship types exist among entities (documents, images, video, etc.). From the small-group discussions, it's also clear that these relationships could be useful in such applications as user profiles, intelligence analysis, categorization, etc.

A number of challenges lie ahead to enable systems to take advantage of this kind of document information, including new methods for extracting relationships, mathematics for representing them, and visualizations allowing users to explore and understand them.

When asked what were the most important areas for follow-on research, there were two areas that were consistently emphasized by the participants:

- developing an understanding of how the various kinds of document relationships are important to support the variety of user needs and tasks
- developing the mathematical methods for representing the relationships

Several workshop participants are actively exploring this concept of non-traditional relationships among documents or other information objects. It is hoped that this workshop and summary have taken us at least a small step further along the path toward understanding and addressing this important area of research.

A more complete summary of the relationships identified is available. For more information on this workshop, questions about the article, or suggestions for follow-on activities, contact Beth Hetzler at beth.hetzler@pnl.gov.

References:

Eick, S. G. March 1996. "Aspects of Network Visualization." *IEEE Computer Graphics and Applications*. 69-72.

Fidel, R. and Crandall, M. 1997. "Users' Perception of the Performance of a Filtering System." *In Proceedings of ACM SIGIR '97*, pp. 198-205. ACM Press, New York, New York.

P1000 Planning Committee. 1996. "Visualization of Structured Data Sets." *P1000 Science and Technology Strategy for Information Visualization*. 36-44.

Swanson, D. R. January 1990. "Medical Literature as a Potential Source of New Knowledge." *Bulletin of Medical Library Association* 78(1):29-37.