fw consulting

document visualization

introduction

The goal of document visualization is to display documents and document collections so that it is easy to quickly find relevant information. To achieve this, document visualization uses methods that use the capabilities of the human visual system to rapidly absorb and understand information.

Document visualization can help a user to quickly understand what a document is about and where its important passages are. This is particularly useful for long documents. Document visualization also illustrates relations between documents (e.g. how documents cite each other; documents with common topics; documents written by the same authors or at the same place or time). Furthermore, document visualization can help analyze temporal patterns in document collections (e.g. when coverage of certain topics in- or decreases).

Here we briefly illustrate some methods for document visualization. All these methods are active areas of research and development at FW Consulting.

discover topics of documents and document collections

One way to quickly see what a document is about is to identify and present its topics. A topic indicates what a cluster of related words or phrases from a document is about. Presenting a word cluster with a topic label is much more informative than just presenting the word cluster itself. At FW Consulting we are developing methods for clustering related words or phrases in a document, as well as methods for automatically discovering topic labels for these clusters.

The constructed example in Figure 1 illustrates how a topic label makes a word cluster more informative. The cluster on the left shows words from a document (e.g. a scientific paper). These words are somehow related in the document (e.g. they appear in close proximity of each other). However, this cluster of words has relatively little informational value by itself. All one can tell is that the strings are probably last names. By contrast, the cluster on the right is much more informative. The topic label "previous research" suggests that the strings are last names of researchers whose work is discussed in the document.

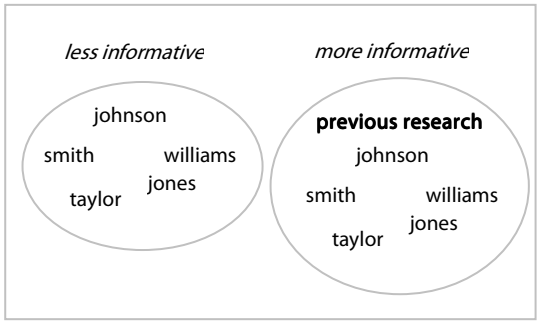

Figure 1. Cluster labels can make clusters more informative.

For individual documents, topics help a user get an idea of what the document is about. But topic discovery can also be applied to document collections. In the case of document collections, topics are used as labels for clusters of related documents. Topic discovery thus helps categorize the documents in the collection. Of course, this is particularly useful if it is not known in advance what the categories or topic labels are (e.g. that some documents are about sports, some about finance, and some about cultural events).

understand relations between documents

There are several ways in which documents in a document collection can be related. For example, documents can cite each other, discuss common topics, belong to the same genre (e.g. scientific paper vs. newspaper article), or they can be written by the same author, at the same time, or at the same place.

A graph like in Figure 2 illustrates how documents cite each other. The arrows point from the document that cites to the document that is being cited. Figure 2 shows that document 4 is the most-cited document. The fact that a document is cited frequently often indicates that this document contains important

information. Examples include an influential scientific finding, an important event, or a well-known author.

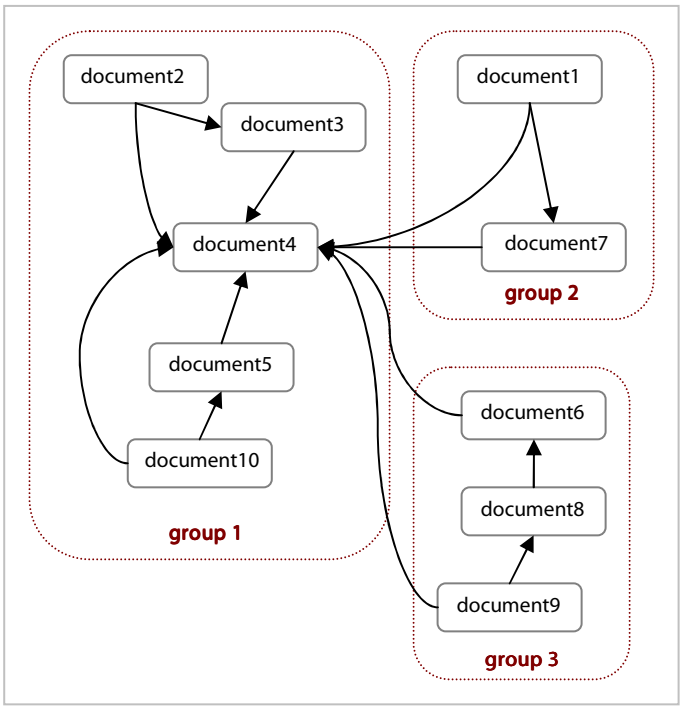

Figure 2. Citation networks show frequently cited documents and citation groups.

Figure 2 also shows that there are three citation groups, indicated by the red rectangles: group 1 consists of documents 2, 3, 4, 5, and 10; group 2 includes documents 1 and 7; group 3 consists of documents 6, 8, and 9. Groups 2 and 3 are related to group 1 via citations of document 4. Citation groups could indicate common topics or common authors, for example.

Of course, the constructed graph in Figure 2 is just a tiny example. Real-world graphs are usually orders of magnitudes larger (i.e. they contain many more documents and many more citation links). This makes finding citation groups a lot harder and requires advanced computational methods. At FW Consulting we are developing such methods.

As mentioned above, documents can also be related in other ways. For example, they can have common topics, belong to the same genre (e.g. scientific paper, newspaper article, technical report, or comment), have the same author, or they can be written at the same time or place. Such commonalities can be visualized in a heatmap matrix such as in Figure 3. "Topic" in Figure 3 could also be "genre", "author", "time", or "place", for example. The shading of the cells indicates how strongly a topic is represented in a document: darker shading indicates stronger representation.

Figure 3 shows that document 5 covers all topics (although topic 1 is covered a bit less than topics 2 and 3, as indicated by the cell shading). On the other hand, for example, document 9 only covers topic 1. Figure 3 also shows that topic 2 is covered most and that topic 1 is covered least.

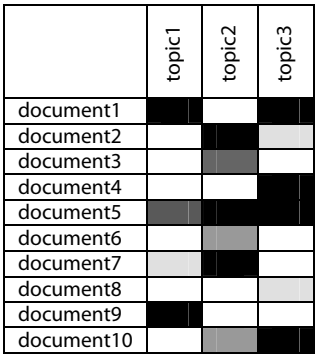

Figure 3. Heatmap matrix showing commonalities between documents.

Heatmap matrices can be made much bigger than the constructed example in Figure 3 (i.e. have more rows and columns), and still be intuitive and fast to understand. Thus, heatmap matrices exploit the capability of the human visual system to absorb and understand great amounts of information simultaneously.

see whether documents are relevant to a search query

In traditional search engines, the user first enters one or more search keywords or phrases (the *search query*). The search engine then returns list of relevant documents, ranked by how closely related they are to the search query. The point of the ranked list is to help the user decide quickly which document they should look at more closely.

A ranked list of relevant documents usually contains the title of each document and a one- or two-line example passage from each document where the search term appears. Many search engines also mention some figure of relevance for each document. For example, a highly ranked document might be 89% relevant, whereas a less highly ranked document might only be 63% relevant. These relevance figures are meant to indicate how strong the match is between a document and the search query.

However, there are two problems with this traditional approach:

- Displaying only a one- or two-line example of where in the document the search query appears might not be enough. The search query often appears more than once in a document (particularly in long documents). Some appearances of a search query might be more informative than others.
- It is unclear what the relevance figures really mean. It is not transparent to the user what exactly the difference is between "89% relevant" and "94% relevant", for example.

A simple graphical display as shown in Figure 4 can address these problems[1]. A user can then see at a glance if it is worth taking a closer look at the document itself. In Figure 4, the length of the bars represents document length (i.e.

[1] Similar displays have also been developed by Dr. Marti Hearst at Berkeley.

document 3 is the longest of the three documents). The lines represent places in the document where the search query appears. The darkness of the lines represents the prominence of the search queries in the places where they were found; prominence could be frequency or whether the query appears in a structurally salient part of the document (e.g. in a section heading). As Figure 4 shows, document 1 has the most matches of the search query.

Figure 4. Display showing the location and prominence of search queries in a document.

These displays can also be used to visualize how strong the match is between a document on one hand and several queries on the other hand. For example, Figure 5 shows that queries 1 and 2 occur several times in the document, but query 3 occurs much less frequently (i.e. the match between the document and queries 1 and 2 is stronger than the match between the document and query 3). Furthermore, Figure 5 also shows that queries 1 and 2 occur in very similar locations in the document.

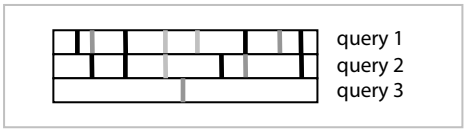

Figure 5. Display showing how different search queries are represented in one document.

highlight important passages in a document

One approach to finding important passages in a document is to extract these passages and then present them to a user. An alternative approach is to present the whole document and just highlight the important passages. This is equivalent to what people do with paper documents: they mark important passages with a highlighter pen (e.g. in yellow).

The advantage of highlighting important passages instead of extracting them is that the user can see the passages in their original context. This exploits another capability of the human visual system, parafoveal vision: the capability to absorb and understand information that is presented outside the current focus of attention. For example, a user might focus their attention on a highlighted passage, but they are still able to understand material surrounding the highlighted passage.

discover temporal patterns in document collections

The topics that are covered in document collections often change over time. Newswires are examples of such document collections. For example, if there is a big sports event, such as Olympic Games, newswire messages about sports will probably increase. If there is a political crisis in the Middle East, there might be an increase not only of messages about the Middle East. There might also be more messages about the oil market, or about renewable sources of energy.

Figure 6 shows a constructed example of how the topics in news article collections might change over time[2]. In 2006, the Olympic Winter Games could be responsible for more reports on downhill skiing, biathlon and figure skating. The Australian Open, French Open, and Wimbledon tennis tournaments are probably the reason for the peaks in the numbers of tennis reports. The peak in the number of soccer reports in June and July could be due to the Soccer World Cup.

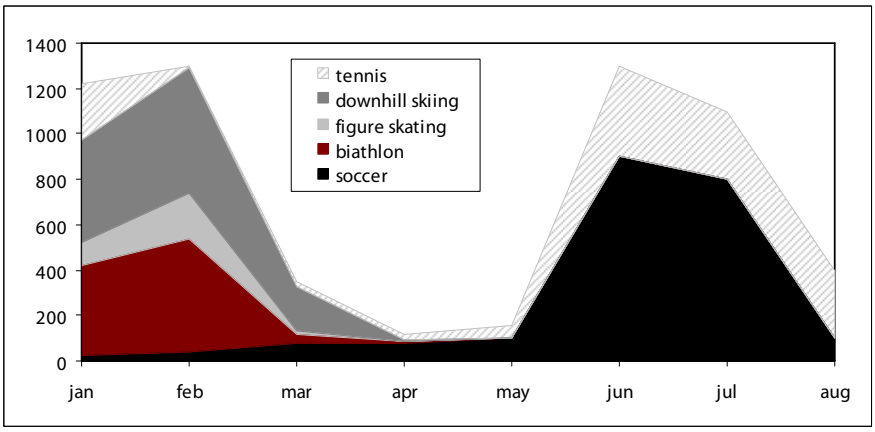

Figure 6. Development of topics over time in a collection of documents.

Displays like in Figure 6 can also help discover relations between different topics. For example, if the coverage of two topics increases at the same time, these topics might be related. An example in Figure 6 is the coverage of downhill skiing, biathlon, and figure skating. Coverage of all these topics increases at around the same time. The Olympic Winter Games are the reason for this increase.

Consider another example of how displays like in Figure 6 can help form hypotheses about events or topics: Imagine a simultaneous increase in newswire messages covering the topics "oil price", "oil industry" and "Saudi Arabia". A possible hypothesis, based on this pattern, is that some action or event in Saudi Arabia had some impact on the oil market (because Saudi Arabia is a very important oil producer). This hypothesis could not have been formed based only on the temporal patterns of the topics "oil price" and "oil industry" alone.

[2] Similar displays have also been developed by Dr. Susan Havre and colleagues at the Pacific National Laboratory.

technologies required for document visualization

Most of the displays for document visualization that were illustrated here are relatively simple. But to be useful, these displays require a number of sophisticated methods and technologies "under the hood". These methods include topic detection and tracking, graph clustering, automatic construction of lexicons, and information extraction from text. All of these methods are active areas of research and development at FW Consulting.

contact

postal address	Dr. Florian Wolf Am Mitterfeld 34 85570 Ottenhofen	fon fax email www	+49 (0)8121 22 76 68 +49 (0)8121 22 76 69 wolf@f-w-consulting.com http://www.f-w-consulting.com