Towards Retrieving Relevant Information Graphics

Zhuo Li
ivanka@udel.edu

Matthew Stagitis
mattstag@udel.edu

Sandra Carberry
carberry@udel.edu

Kathleen F. McCoy
mccoy@udel.edu
Department of Computer and Information Science
University of Delaware, Newark, DE 19716

ABSTRACT

Information retrieval research has made significant progress in the retrieval of text documents and images. However, relatively little attention has been given to the retrieval of information graphics (non-pictorial images such as bar charts and line graphs) despite their proliferation in popular media such as newspapers and magazines. Our goal is to build a system for retrieving bar charts and line graphs that reasons about the content of the graphic itself in deciding its relevance to the user query. This paper presents the first steps toward such a system, with a focus on identifying the category of intended message of potentially relevant bar charts and line graphs. Our learned model achieves accuracy higher than 80% on a corpus of collected user queries.

Categories and Subject Descriptors

H.3.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Information Search and Retrieval, Digital Libraries

Keywords

Machine Learning; Natural Language Processing; Query Processing; Graph Retrieval

1. INTRODUCTION

Research on information retrieval, information extraction, and question answering have focused almost exclusively on information available from text and, to some extent, from images. Information graphics (non-pictorial images such as bar charts and line graphs) have been largely ignored. Yet such graphics contain a wealth of information that should be easily accessible.

Document retrieval relies on matching words in a query with words in documents, using expansion of queries with related words and metrics such as tf-idf. When retrieving information graphics, search engines, such as Google Image and Zanran, both rely heavily on the text from the source file that contains the information graphic, including the graph's file name, the image tag from the webpage html source file, or words in the accompanying article that appear near the graphic's geographical location in the article. This approach does not take into account the content of the information graphic itself. For example, when the query *"How does the revenue of Google compare with the revenue of other technology companies?"* was entered into Google Image search on Feb 20, 2013, the highest relevancy ranked graphics returned were off-target. The graphic deemed most relevant was a bar chart showing the revenue of the iTunes App Store, Amazon AppStore, and Google Play; this graphic appears to have been selected since the words *Google*, *technology*, and *revenue* appear near the graphic in the accompanying article. However, this graphic is much less relevant than the graphic in Figure 1 which was also available.

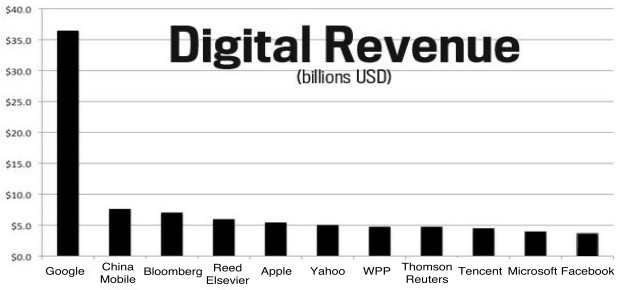

Figure 1: Revenue of Technology Companies

We contend that retrieval of information graphics should take into account how the informational needs of the user are satisfied by the content and structure of candidate graphics. For example, Consider the following two queries:

Q_1: *Which countries have the highest occurrence of rare diseases?*
Q_2: *Which rare diseases occur in the most countries?*

These two queries contain almost identical words but are asking for completely different graphics. Query Q_1 is asking for a comparison of countries (independent axis) according to their occurrence of rare diseases (dependent axis), while query Q_2 is asking for a comparison of different rare diseases (independent axis) according to the number of countries in which they occur (dependent axis). The difference between these two queries cannot be determined by keyword match-

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.

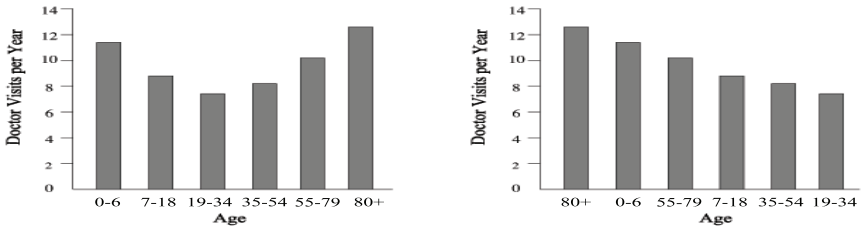

Figure 2: Two Graphs Displaying the Same Data

ing alone.

Similarly, consider the following query:

Q_3: How does the number of doctor visits per year change according to a person's age?

Although both of the graphics in Figure 2 contain the same data, the changing trend in doctor visits by age is more easily seen in the leftmost graphic than in the graphic on the right. Instead of reversing the axes as was done in the previous example, these two graphics differ in the high-level message that they are intended to convey. This correlates with an observation by Larkin and Simon[8] that graphs may be informationally equivalent (that is, they contain the same information) but not computationally equivalent (that is, it may be more difficult for humans to extract certain pieces of information from one graphic than from the other).

Our research is concerned with the retrieval of bar charts and line graphs that appear in popular media. Because of the sparsity of words in a graphic and the importance of a graphic's structure, retrieval cannot be done on the basis of simple metrics such as tf-idf as might work for textual documents. A deeper analysis of graphics and their relationship to the user's informational needs is required. But accomplishing this requires that the user's query contain more than just a set of keywords. Thus we are developing a graph retrieval system where the input is full sentence question queries whose semantics can be analyzed to identify characteristics of relevant graphs. Our current work is focused on the first steps toward such a system: extracting from the user's query the content of the dependent and independent axes and, the focus of this paper, the category of the intended message of potentially relevant simple bar charts and single line graphs.

2. RELATED WORK

State of the art content-based image retrieval has been making progress in judging semantic similarity from visual similarity [2]. Some of the image retrieval systems rely primarily on the text from the multimedia document [7]. Some image retrieval systems use automatic annotation learned from a manual annotation set [6], or user-provided metadata tags in social media such as Flickr and Youtube [5]. Research on information graphics has focused on classification of the type of graphic (such as bar charts or line graphs) [10], or information extraction by converting information graphics into tabular form data [9]. To the best of our knowledge, our work is so far the only project which attempts to recognize the high-level message or knowledge conveyed by bar charts and line graphs and use it in determining the relevance of a graphic to a user query.

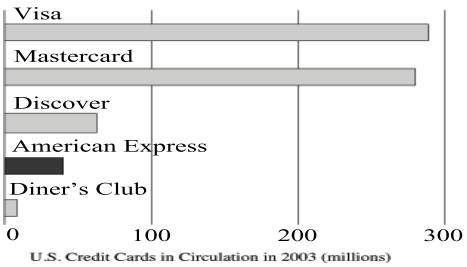

Figure 3: Graph with a Rank Message

3. UNDERSTANDING THE CONTENT OF INFORMATION GRAPHICS

Information graphics that appear in popular media such as magazines and newspapers generally have a high-level message that they are intended to convey. For example, the leftmost graphic in Figure 2 conveys the changing trend in doctor visits over one's lifetime whereas the rightmost graphic conveys the rank of different age groups in terms of number of doctor visits.

We identified a set of message categories that capture the kind of messages that might be conveyed by simple bar charts and single line graphs. These consist of: *Trend* message categories that convey a trend over some ordinal entity, comparison message categories such as *Relative-difference* that contrasts two entities, *Rank* that conveys the rank of an entity with respect to other entities, *Rank-all* that conveys the relative rank of a set of entities, and *Maximum* and *Minimum* that convey the entity that is the largest or smallest with respect to some criteria.

We developed systems for recognizing the primary message conveyed by two categories of information graphics: simple bar charts and single line graphs [4, 11]. These systems rely on the presence of communicative signals in the graphic. For example, salience of an entity in a graphic might be conveyed by coloring the bar differently from other bars (as in Figure 3) or by the fact that it is much taller than the other bars (such as the bar associated with Google in Figure 1), thereby suggesting that it plays a major role in the graphic's high-level message. These communicative signals are entered as evidence in a Bayesian network that hypothesizes the graphic's intended message — both the category of intended message and its parameters. For example, the intended message for the graphic in Figure 3 would be formally represented as Rank(American Express, {Visa, Mastercard, ... , Diners Club}), indicating that it is conveying a Rank message and that American Express is the focused entity being ranked against the other listed companies.

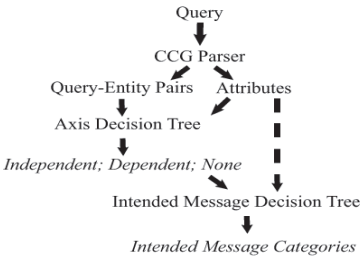

Figure 4: System Overview

Unfortunately, information graphics often do not explicitly label the dependent axis with what is being measured [3]. For example, *digital revenue* is being measured on the dependent axis by the graphic in Figure 1 but the dependent axis is unlabeled. We developed a methodology for hypothesizing what is being measured by the dependent axis of a graphic; this methodology utilizes a set of heuristics that extract information from the dependent axis, from within the graphic itself (for example, the words *Digital Revenue* that appear within the graphic in Figure 1), from the main caption on the graphic, from any elaboration of the main caption, and from the caption on a composite of several graphs, and melds it together to form a measurement axis descriptor.

In this paper, we assume that each graphic is stored with its intended message, the labels on its independent axis, and the measurement axis descriptor capturing the content of its dependent axis.

4. METHODOLOGY FOR HYPOTHESIZING MESSAGE CATEGORY

Our methodology is to extract clues from the user's query and use these to learn models for identifying the content of the independent and dependent axes and the category of intended message of potentially relevant graphs. Figure 4 outlines the application of the learned model to hypothesizing the content of the axes and intended message of potentially relevant graphs.

Given a new query, the system first passes it to a CCG parser [1] that is trained especially for questions to produce a parse tree. From the parse tree, we populate a set of candidate entities E_1, E_2, . . . , E_n that might capture the content of one of the axes, and extract a set of linguistic attributes associated with each query-entity pair Q-E_i. Two examples of such attributes are:

- Type of query: *Which* and *What* queries are often followed by a noun phrase that indicates the class of entity (such as *countries*) that should appear on the independent axis, whereas *How much* and *How many* queries are often followed by a noun phrase that indicates what quantity should be measured on the dependent axis.
- Superlative/comparative: The presence of a superlative or comparative, such as *"highest"* or *"higher"*, often suggests that the dependent axis should measure the noun phrase following the superlative or comparative.

The attributes for each query-entity pair are input to a decision tree for determining whether the entity represents the content of the independent axis or dependent axis. Then we use the content of the axes to help identify the category of intended message requested by the user's query.

The content of the axes, as identified from the query, plays a significant role in identifying the category of intended message of graphics that might be relevant to the query. For example, if the query indicates that the independent axis should represent a time interval, then the intended message of relevant graphics is likely to fall into the trend category. Similarly, the number of entities depicted on the independent axis is a clue about the intended message of relevant graphics. Consider the following example queries:

Q_4: *How does the revenue of Google compare with that of other technology companies?*
Q_5: *How does the revenue of Google compare with that of Facebook?*

Knowing that *Google* and *technology companies* are components of the independent axis, and that one is singular (*Google*) while the other is plural (*all technology companies*), suggests that query Q_4 might be asking for a graphic whose intended message falls into the *Rank* category, namely the rank of *Google* among all technology companies. On the other hand, knowing that *Google* and *Facebook* are the entities on the independent axis and that both are singular suggests that query Q_5 might be asking for a graphic whose intended message category is *Relative-difference*, namely a comparison between *Google* and *Facebook*. Although an information graphic that includes the revenue of many technology companies, including Google and Facebook, could provide the information requested by query Q_5, the user can extract that information more easily from a graphic specifically devoted to Google and Facebook without other entities to distract the reader's attention. Thus attributes based on the identified content of the axes include the number of independent axis entities and their plurality.

The class of the main verb in the user's query is also useful in hypothesizing the intended message of relevant graphs. For example, *comparison* main verbs, such as *differ* and *compare*, suggest that relevant graphics will have a *Relative-difference* or *Rank* intended message; on the other hand, main verbs in the *change* class suggest a trend message. Superlatives, such as the word *highest*, suggest that relevant graphics will have a *Maximum* or *Minimum* intended message. Space prevents discussing all of the attributes used.

5. EVALUATING THE METHODOLOGY

We conducted two human subject experiments to construct a corpus of full-sentence queries oriented toward retrieving information graphics. Each subject was shown a set of information graphics on a variety of subjects. In the first experiment, for each displayed graphic, the subject was asked to construct a query that could be best answered by the displayed graphic. In the second experiment, each participant was given several sets of information graphics; each set consisted of four graphs with similar data but different intended messages. For each graph in a set, the subjects were asked to write a query where that graph would be more relevant than the other graphs in the set. The two experiments produced a corpus of 324 queries in total.[1]

[1]The links to the online SQL databases are www.eecis.udel.edu/∼stagitis/ViewAll.php, www.eecis.udel.edu/∼stagitis/SE/ViewAllSets_Graphics.php

Given a query, the system must identify the category of intended message for potentially relevant graphics. Our current work has been limited thus far to simple bar charts and single line graphs, and the categories of intended messages are *Rank*, *Rank-all*, *Relative-difference*, *Min-max-single*, *Min-max-multiple*, *Trend*, *General-single*, and *General-multiple*. The *General* intended message category (*General-single* and *General-multiple*) were added to capture situations where the query was not specific about a particular category of intended message; an example is the query *"How many millions of daft punk albums were sold in 2000?"*. The *Single* versus *Multiple* distinction captures the difference between requesting a single entity (*"What is the GDP of the U.S.?"*) versus multiple entities (*"What is the GDP of various developed countries?"*).

Models were learned for hypothesizing the content of the axes and the category of intended message of potentially relevant graphs. Using leave-one-out cross validation, our system had a success rate of 81.48%, which is much higher than the baseline of 58.33% which is achieved by simply selecting the most prevalent category (namely *Trend*).

6. NEXT STEPS FOR GRAPH RETRIEVAL

We are developing a mixture model that will use a variety of features in retrieving the most relevant graphs in response to a full-sentence user query. By identifying the appropriate content of the independent and dependent axes, we will be able to reduce the number of graphs that need to be considered. By identifying the intended message specified by the query, we will be able to both eliminate and rank graphs for retrieval. For example, if the message category identified from the query is *Rank*, then graphs whose intended message category is *Trend* can be eliminated. A hierarchy of message categories will be used to relax the required message category when appropriate. For example, although a graph with a *Rank* intended message that draws attention to the entity being ranked (such as the graph in Figure 3) would be most appropriate in response to the query *"How does the number of credit cards for American Express compare with that of other credit card companies?"*, a graphic with a *Rank-all* message that conveys the rank of all credit card companies (without drawing attention to any single company) would be acceptable but less desirable.

An issue that must be addressed is the indexing of graphics in a digital library. Documents are indexed by the words in the document, but there are few words in a graphic. Since it would be time-consuming to evaluate the relevance of every available graphic to the user's query, we need a means of selecting a subset of the graphics for consideration. Although not discussed in this paper, we are using Wikipedia to expand the terms in a graphic and produce an expanded set of words that will be used to index the graphic. Given a user query, the words in the query will be used to preselect a set of candidate graphs. For example, if the labels on the independent axis of a graph are *Google*, *Yahoo*, and *Apple*, our approach will produce an expanded set of words that includes *technology* and *company*, thereby allowing us to select this graph as a candidate in response to a query such as *"Which technology companies are most successful?"*.

7. CONCLUSION

This paper has presented the first steps in the development of a system for effectively retrieving information graphics in response to a user query. Our method relies on full-sentence queries in order to identify features of potentially relevant graphics, rather than relying merely on keyword matching. Thus far, we have developed learned models for identifying the content of the independent and dependent axes and the category of intended message of relevant graphs. Future work will utilize these in a mixture model for ranking graphs for retrieval. To our knowledge, this work is the only research effort that is specifically focused on the retrieval of information graphics and that is attempting to take into account the content of graphics.

8. ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation under grant III-1016916.

9. REFERENCES

[1] S. Clark and J. Curran. Wide-coverage efficient statistical parsing with ccg and log-linear models. *Computational Linguistics*, 33(4):493–552, 2007.

[2] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (CSUR)*, 40(2):5, 2008.

[3] S. Demir, S. Carberry, and S. Elzer. Effectively realizing the inferred message of an information graphic. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 150–156, 2007.

[4] S. Elzer, S. Carberry, and I. Zukerman. The automated understanding of simple bar charts. *Artificial Intelligence*, 175(2):526–555, 2011.

[5] Y. Gao, M. Wang, H. Luan, J. Shen, S. Yan, and D. Tao. Tag-based social image search with visual-text joint hypergraph learning. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 1517–1520. ACM, 2011.

[6] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 119–126. ACM, 2003.

[7] M. Lapata. Image and natural language processing for multimedia information retrieval. *Advances in Information Retrieval*, pages 12–12, 2010.

[8] J. H. Larkin and H. A. Simon. Why a diagram is (sometimes) worth ten thousand words. *Cognitive science*, 11(1):65–100, 1987.

[9] A. Mishchenko and N. Vassilieva. Chart image understanding and numerical data extraction. In *Digital Information Management (ICDIM), 2011 Sixth International Conference on*, pages 115–120. IEEE, 2011.

[10] M. Shao and R. Futrelle. Recognition and classification of figures in pdf documents. *Graphics Recognition. Ten Years Review and Future Perspectives*, pages 231–242, 2006.

[11] P. Wu, S. Carberry, S. Elzer, and D. Chester. Recognizing the intended message of line graphs. *Diagrammatic Representation and Inference*, pages 220–234, 2010.