Mining Journal Article Components: the Case of Map Graphics

Judith Gelernter
Language Technologies Institute
Carnegie Mellon University, U.S.A.

Categories and Subject Descriptors
H.3.1 **[Information storage and retrieval**]: Content Analysis and Indexing—
Indexing Methods, *Thesauruses*

General Terms
Algorithms

Keywords
text mining, metadata harvesting, information extraction, indexing, information retrieval, graphics, theme maps

Abstract

Finding diagrams, chemical structures, maps or other journal article components is difficult because the articles typically are indexed by topic. Many of these components are all but invisible to standard keyword search. This research describes a method to extract component maps along with metadata from journal articles, and a way to index the maps to make them retrievable by category browse as well as keyword search. Proposed ranking of retrieved maps is by semantic relevance to the query as well as by map attributes such as scale or resolution. The method for extraction and indexing of maps with only minor adjustments should be generalizable to other sorts of article components.

1 Introduction

Participants in a user study asserted that there is a "consistent, unmet need for systems that yield higher precision searches…[to] journal article components like figures, tables, graphs, maps and photographs" [Sandusky & Tenopir, 2008, p.977]. An earlier study of an article component database found that some users wished to retrieve tables and figures before viewing the entire article [Bishop, 1998].

We cannot presently find journal article components easily because articles are indexed by main topic, which is not necessarily the topic of the component graphic. We know from study findings that keyword search can be inadequate to find components [Carberry et al., 2006]. The researchers analyzed 100 randomly-chosen graphics to determine whether the information contained in the graphic was fully conveyed by the text. In only 22% of instances did the document fully convey the graphics' message, in 17% the document mostly conveyed the message, in 26% the document conveyed little of the message and in 35% of the cases the document conveyed none of the graphic's message [ibid, Table 1]). We conclude that in about 35% of cases, journal article components will be all but impossible to find by standard search.

Another approach is to aggregate components in a database that optimizes searches for them, with the option to view the full source if desired. For example, a database of software has been collected in which users look for sections of code that can be re-used in another software [Yao et. al, 2008]. A database of definitions and summaries and such that can be used for teaching has been set up to be searched separately from the slide presentations in which they are found [Verbert et al., 2005]. Biomedical images also have been collected for independent search [Liu et al., 2005]. Google Images is a means of independently viewing web page components. In future, some maps embedded in web pages will be retrievable with the new W3C-condoned Scalable Vector Graphics language that describes two-dimensional graphics. The language is new enough that it is not standard at this writing, however, and retrieval of graphics is not always efficient.

This research focuses on article components which are maps. Even with the existence of geo-libraries to hold maps, many look for maps by means of a general search engine. Queries on geography contributed to the largest percentage of queries in any category to a large online search engine [Sanderson & Han, 2007]. (The other categories were coded as activity, adult, arts and humanities, shopping, computer, education, healthcare, people and science.)

Are people finding what they look for? An experiment conducted in 2001 with high school and college freshmen and repeated in 2005 with college juniors, showed that for many, finding a continent map was easy, but finding a .pdf version of a map was more challenging [Peterson, 2006]. Maps from journal articles will oftentimes be native in .pdf format because this is a preferred storage format for the articles that contain such maps. Journal articles are fertile sources to look for maps. Advantages of taking maps from journals include that the article could be a source of additional information on the theme, journal title could be used for clustering, and maps could be ranked by number of citations of the article.

Our focus is not on maps of position, as exemplified in local area search, business search (finding where a certain business is located), and driving direction and route search. Local and regional government websites supply maps, as do tourist websites. Some such are actual cartographic representations used for reference or for navigation purposes. Nor is our focus on printed historical or reference maps that have been digitized. Nor is our focus on newly-made map mashups in web 2.0 applications such as Platial or GeoCommons.

Instead, our focus is on *thematic maps*. These use a geographic base to show the distribution of a theme. Motivation stems from a double need: to mine journal articles for components, and to find theme maps.

The need becomes greater when you consider presently how to find such a map. You could create the map yourself, but that requires knowing sites associated with the mapped theme. Instead, you could take an obvious approach and search the web for what is available. What if you seek, for example, a map of different forms of wildlife and their habitats around Antarctica? You could consult online wildlife journals, although you might have to spend a long time browsing many articles before you find something satisfactory. Alternatively, you might turn to Google Images, but you would probably be disappointed by

the results. At the time of writing, the top 20 results include not a single map labeled with wildlife (although one retrieved image shows an ice floe with frolicking whale's tail. Only 14 of the images retrieved show the entire continent. Five show parts of the continents in larger scale. One result retrieved shows travel routes of ships along which one can view wildlife. In brief: none of the retrieved images show a distribution of wildlife in general or any species in particular over Antarctica, and there are many false positives. There is work to be done in this area.

This research is in response to what we see as the need to find theme maps. Questions guiding research include:

- How can we perform efficient retrieval for article components (in this case, maps)?
- What information do we need to extract in order to improve image retrieval?
- How could we use ontologies to improve map retrieval by region, time period and subject?

Our work does not concern distinguishing what journal article components should be mined. Recognition of the component based on image attributes (chart, diagram, map, for example) has been studied by Lu, Mitra, Wang and Giles [2006] and by Tan, Mitra and Giles [2009]. A short description of image extraction is found in Gelernter and Lesk [2009]. For the purposes of our work, maps were identified manually and clipped from articles using the Adobe snapshot tool and, using Microsoft Paint, stored in .jpg compression format. Our work also does not consider an optimal approach to scan the image so as to retrieve the map title and labels, although these words are clearly valuable for retrieval. The words-in-map, too, were mined manually for this research.

Below §2 describes related work and justifies our choice of method. §3 presents specifics on the data we used and how we mined metadata for indexing, while §4 discusses classification and browse categories and §5 concerns how ontologies aid the classifiers. §6 describes evaluation procedures, with suggestions for future improvements. The conclusion is in §7.

2 Related Work

This section describes approaches to indexing methods for maps and other graphics. Others have attacked the problem by image recognition (content-based image retrieval (CBIR)), recognizing relationships within the image map (geospatial analysis), recognizing words within the image (document image analysis), or mining the image caption or file name (semantic retrieval).

Image recognition. Retrieval of a graphic based on similarity to others is known as content based image retrieval. Image retrieval has been successful with low-level features such as color [Pass et al., 1996], shape [Syeda-Mahmood, 1996], and edge matching [Liu et al, 2005]. Similar methods have been applied to diagrams [Futrelle & Nikolakis, 1995]. The "Find similar" command of Google Image Search presumably works along these lines. Google Image Swirl not only finds similar images but also clusters them into subcategories (such that a query on "Eiffel tower," for example, finds images during the day, at night, or from a distance).[1] This approach seems inappropriate for theme maps because edge matching might align borders of geographical regions with less attention to superimposed theme points.

Geospatial queries. Closer integration of web search with geographic information systems would allow the extraction of geographical information from local map search systems [Tezuka et al., 2006]. A novel approach to indexing maps uses shapes to retrieve similar shapes within the image. The maps are

[1] http://googleresearch.blogspot.com/2009/11/explore-images-with-google-image-swirl.html

deconstructed into polygons and the query input is another sort of polygon [Zhang et al., 2009]. Geospatial queries seem ineffective for retrieving image-based theme maps of the current study because such maps are more visualizations than cartographic representations, so that the geographic relationships are in many cases secondary to the theme data.

Document image analysis. Extracting words from titles or labels in the map image is a form of document image analysis. A method for the separation of a map into constituent layers in order to analyze the symbols has been developed by Dhar and Chanda [2006]. Complicated techniques have been implicated to extract text contained in color images [Chen, 2008], but others have found a light weight optical character recognition scan of an image adequate to extract text [Hirano et al., 2007]. Our own preliminary comparisons of the open source OCR program Tesseract with ABBYY's FineReader and LizardTech's Document Express suggested that Document Express is best for extracting words from the map. However, this technique finds few labels over geographic areas, because the contrast is usually not good enough to read words.

Semantics of article or filename. Mining words under an image has been used effectively to index image libraries [Guglielmo & Rowe, 1996]. It has been proposed that captions be used to retrieve images, and even video and audio files [Rowe, 2002]. Another method uses the file name to determine content of the graphic [Li et al., 2005]. Although this can be successful on a more general level, the file name does not present much information on what the image contains and is, of course, inaccessible in a .pdf article.

In light of the possibilities outlined in this section, we advocate use of words for the retrieval of map graphics. We show below that the caption, as well as related sources of metadata such as the title of the article and any sentence in the article that refers to the map will be useful for information retrieval. Further, words in the map itself aid map retrieval.

3 Data and Data Mining

We combed articles manually from many disciplines in order to find 150 maps. A few maps were from the same article, but most were from different articles in different journals. The maps differed also in format: large and small, color and black and white, small scale and large scale. Most maps were extracted in .pdf format since they came from .pdf articles; a few were from web pages. All were converted to .jpg for manipulation in the prototype database. The maps were considered only if they had a minimum resolution of about 200 x 300 pixels so as to be clear for on-screen viewing.

Questions guiding data extraction research were what metadata to mine (that would be most useful for information retrieval), how to prioritize what was mined, and how to actually do the mining.

In considering what metadata to mine for map retrieval, a pressing question is one of extent: how much is enough? A balance must be struck between mining sufficient metadata to classify each item precisely, and too much metadata that will let in noise and cause recall to suffer. The amount of metadata depends on the image type. A typical figure in biomedical articles, for example, may comprise multiple panels, and the text in the image might include names such as proteins. Determining which words best explicate biomedical images has been done effectively by probabilistic methods [Ahmed et al., 2009]. Another approach to mining metadata for biomedical images was to use figure captions plus three relevant sentences in the text that referred to the figure—the one sentence with the direct reference to the figure, and the preceding and following sentence. It was found that figure caption by itself was more effective for classification than the combination of both caption and sentences in the article that referred to the figure [Koike & Takagi, 2009, p. 852].

Others' research directed at maps extracted from documents took for metadata the map caption, sentence in the article that refers to the map, article title, article abstract, and location names extracted from these [Tan et al., 2009]. By contrast, we did not use the article abstract unless other metadata sites yielded nothing. But we included titles and large labels within the map graphic, which we call "words-in-map" as an additional metadata field.

Once the metadata has been collected, how might the terms be prioritized? Tan, Mitra and Giles [2009] used frequency of occurrence in the metadata set and in the collection used a version of TF-IDF they call MTF-MITF, with the "M" standing for map term. Their weighting system ranks higher those maps that come from higher quality documents. They compute document quality from the number of other publications that cite the host document.

We determined priority of the caption, words-in-map, title and referring sentence by manual inspection for good indexing terms. We found the words under the map and the words in the map to be equivalent in indexing potential on the whole. That is, the words within the map image were highly descriptive of map content, just as was the map caption. The article title was helpful to a lesser extent, and the referring sentence was only useful occasionally. Given these patterns, we set weights to make words within and underneath the map worth most, words in the article title worth half of that, and words in the referring sentence a quarter. We arbitrarily assigned numbers to these proportions, below.

Weights for words in the target metadata

Caption words	8 x number of occurrences
Words-in-map	8 x number of occurrences
Title words	4 x number of occurrences
Referring sentence	2 x number of occurrences

How can we actually extract the map caption and sentence that refers to the map? Others have looked at horizontal and vertical placement of the caption with respect to the graphic [Khurram et al., 2009]. While extraction of the map graphic was beyond the bounds of this Natural Language Processing-centered research, rules for extraction of the metadata appear below.[2]

[2] These heuristics appear also in the appendix of Gelernter, *Intelligent Information Retrieval for Maps*, 2009.

A. Caption

1. Locate map.
2. To find map caption, scan directly below and directly above the map for text that is of a different size than the article text, and also scan the side of the map if the map does not take the entire width of the page.
3. Mine data: Take entire caption, either above or below map, from start to end of different size text. (The caption does not necessarily end with a period.)
4. If "Source" or "Reproduced" or "Reprinted" or "©", "Courtesy of" "By permission" or "permission" appears in the caption, do not mine these words or symbols and the text that follows.

B. Title and subtitle of article

1. Go to the beginning of the article for the word(s) larger than the article text
2. Mine entire word string until the word "by"

C. Referring sentence (sentence in the article that refers to the map)

1. The figure number for the article begins the map caption.
   a) Scan the full text to find a match with the indicator. The match could be exact, or with variation: abbreviated with or without punctuation, or possibly in a different case, or with a range of numbers.
   So for example, if the map indicator is Figure 2, search within article for "Figure 2" or "Fig. 2" or "Fig 2" or "fig 2".
   For example, if the map indicator is Fig 2a, search Fig 2a and Fig 2(a) and Fig 2(A).
   For example, if the indicator is "APPENDIX B," search for "Appendix B". If the indicator is "Illustration," search for keyword "Ill." or "Ill" or "Illustration".
   For example, Fig 1 might match with Figs. 1-x, or Figs. 1, or Figures 1, etc.

   b) If step C1a) finds nothing, scan caption for words in pairs with or without intervening stop word. Start at caption beginning and run through the entire caption in pairs.
   For example, look through the full text for [caption word1+caption word2]. If this finds nothing, search for an exact match for non-stop word pair (w2+w3) or (w2 + stop word + w3) in the caption.
   If a keyword match to caption pair is found in the article text, harvest this sentence in lieu of a referring sentence. Unless (see C2 below)

2. Do not mine as specified in steps C1a and Cb if
   a) there is more than one referring sentence. Mine only the first sentence that appears in the article.
   b) exact match is found in a footnote
   c) exact match is found in List of Figures, Table of Figures, Table of Contents, or List of Illustrations
   d) if near exact match such that Fig. 3 matches with Fig. 3.1

Stop harvesting metadata unless C procedures yield nothing, meaning that there is no referring sentence. Then continue to D:

D. Go to abstract or beginning of article

1. Recognize abstract as it is often labeled "Abstract", or else, it might be in a smaller font, or darker font in boldface, or else, simply go to the beginning of the text. Mine up to the first period, if present, or
2. Mine first sentence of first paragraph of the full text if no abstract is present.

Stop harvesting metadata.

4 Classification

We classified the maps into three separate indexes—region, time and subject. Some of the advantages of separate indexes according to Martins, Silva and Andrade [2005] are that subject queries which do not require a geographic referent can be processed efficiently, updates to each index can be handled separately, each index can be optimized separately and different result rankings can be supported.

This particular combination of indexes has precedents. Kemp, Tan and Whalley called it the "space-time-theme composite" [2007, p.84]. Perry, Hakimpour and Sheth propose that space, time and theme should be considered as retrieval elements for a basic web search system [2006]. More important for this application, the region-time-subject browse categories reflect the region, time and subject words people use when they ask for maps. We know this from research related to this work, for which we asked 8 map librarians from the United States and abroad how people generally ask for maps. Each librarian was asked to produce 10 or so representative questions encountered most frequently. 95% of the questions these librarians sent could be coded by region, time and subject [Gelernter & Lesk, 2008]. Because a majority of map questions could be constructed from some combination of region, time and subject facet categories, these are a good combination for the retrieval of maps.

Each facet index in the prototype has its own classifier. The classifiers are made from manually-wrought taxonomies, one for region, another for time, and another for subject. The classifiers all look for matches between metadata and ontology words, although each ranks retrieved items differently. The region ontology uses a gazetteer, and the time and subject ontology use a subset of the Library of Congress Classification Schedule.[3] It ranks metadata and the associated map, but each classifier ranks in a different way. The region classifier ranks by geographical hierarchy (region), which essentially corresponds to scale. The time classifier ranks maps chronologically. The subject classifier ranks by semantic relevance, that is, closeness of metadata term to category query.

Browse categories in the prototype align with each classifier. These categories act as pre-set query terms for already-classified items, so retrieval is fast. We supplied only upper-level categories in the prototype with the assumption that sub-categories could be added later for drill down options. We created the categories to cover the full range of item options rather than to balance the small training collection. So the categories for region include the continents, plus the category "world". Categories for time cover the lifespan of the planet ("Prehistory," "Antiquity," "Middle Ages," "Early Modern," "Modern") rather than the lifespan of the collection (mostly 14th-20th century). Upper level browse categories for subject were compressed from the 18 categories of the Library of Congress Classification system. The reach of the label is not always obvious, so the categories unfurl in the interface for clarification.

We did not wish to stray from the Library of Congress categories, but we recognized that some categories overlap.[4] The category "Medicine," for example, may be seen correctly to be a subset of "Science." To compensate for the overlap, we pre-defined which categories overlapped and allowed items to match partially. For example, a medical document would be assigned a higher score, in the Medicine category than it would be in the Science category, even though it could fit into either and be considered partially correct in either.

---

[3] Library of Congress Classification Schedules, retrieved July 28, 2010 from http://www.loc.gov/catdir/cpso/lcco/

[4] The following interchanges were set up as plausible:
History and Travel with Archaeology and Anthropology; History and Travel with Military; History and Travel with Politics and Law;
Society with Religion and Education;
Commerce and Finance with Politics and Law;
Science with Agriculture with Technology and Transportation; Science with Technology and Transportation; Science with Medicine;
Technology and Transportation with Military; Technology and Transportation with Commerce and Finance

All three facets are supported by a separate ontology. When we consider a browse category to be a query, others have used an ontology similarly [Fan & Li, 2006], [Khan et al., 2004]. In query expansion, terms found in the ontology that are semantically related to the query are used as additional match terms. The consequence is that results retrieved may be more relevant, and recall may increase. Our system also uses this smart search technique.

An item is run through each classifier to produce at least one region, one time, and one subject classification as part of pre-processing. Some items were assigned to more than one subdivision of a category, as determined by weighting. We weighted item categories so that if the second highest score were within 25% of the top-scoring category, the item fit two categories.

5 Ontologies Supporting Classification and Querying

We have explained how the words in the map graphic and the words in the article that pertain to the map, such as caption and sentence that refers to the map, become retrieval metadata. We have explained how the system indexes maps in three categories of region, time and subject, and how each category would be supported by an ontology. When a user enters a browse category or keyword, the system compares query to metadata target. Results are displayed according to semantic similarity by way of the ontologies, or according to other types of relevance such as map size or clarity. The basic functions of the system are diagrammed in Figure 1.

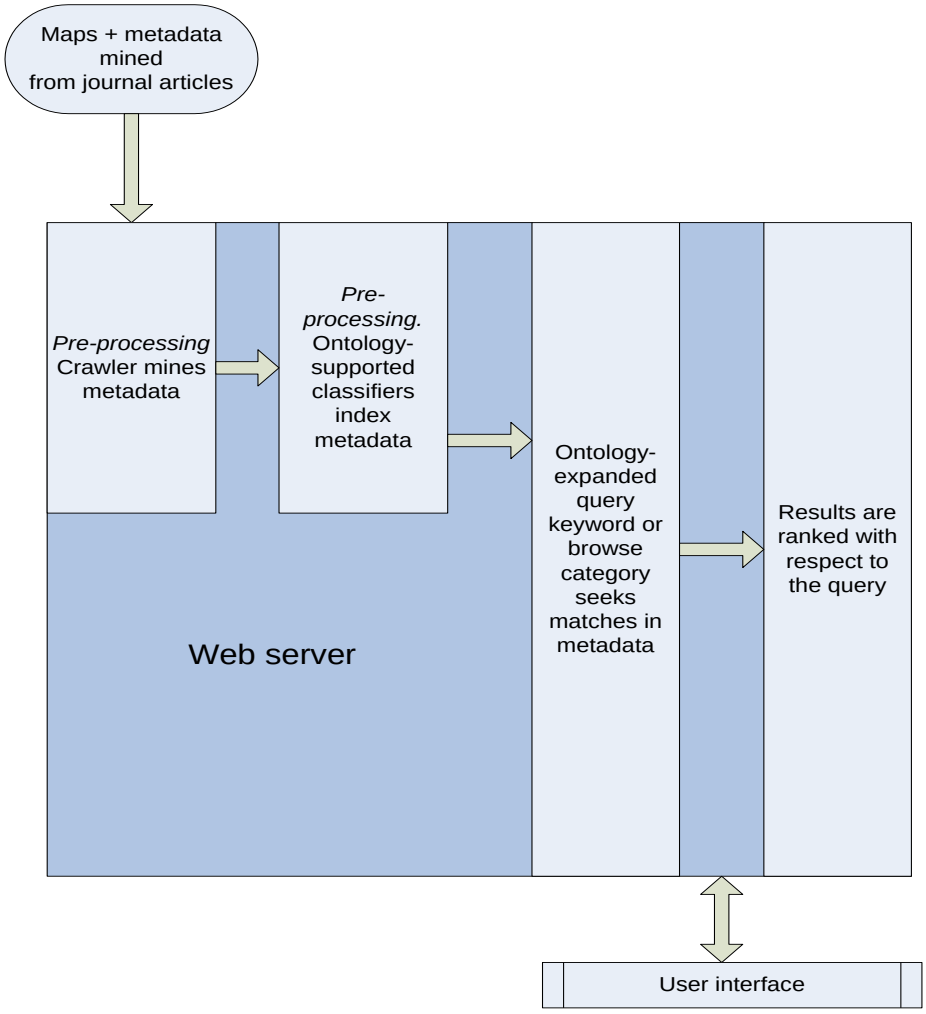

Fig. 1 The architecture of the proposed map retrieval system. As of this writing, a prototype is at http://scilsresx.rutgers.edu/~gelern/maps/.

To clarify, an ontology is an ordered list of terms, sometimes in the form of a hierarchy. It can be used for result ranking as well as to expand a query. Relationships among the ontology terms, when matched with target data, can be used also to show semantic distance among target items for result display. Our system of weighting and ranking is described more fully in Gelernter [2009].

Any of a number of digital gazetteers could have been used for an ontology for region. Leidner [2007], for instance, uses the Geographic Names Information System (GNIS) from the U.S. Geographic Survey. The difficulty with gazetteers for information retrieval is that the more extensive gazetteers include more places that overlap with common words (such as Mobile, Alabama). To reduce false positives, we used the less extensive World Gazetteer for the first pass.[5] Only if we find no place name matches with the World Gazetteer do we run the data through the more extensive GeoNames.[6] The gazetteer is used to list results in order of administrative hierarchy in the gazetteer, which corresponds to the map scale. For

[5] http://world-gazetteer.com
[6] http://www.geonames.org

example, a query for Europe will list first a map of Europe, which is the same scale as the query, and after that, a map of a country in Europe, and at the bottom of the display, maps of European cities.

No working ontology for time is known to us, although temporal information extraction is discussed by Kim and Myaeng [2004] and Mani, Pustejovsky and Sundheim [2004]. Petras, Larson and Buckland [2006] discussed the creation of a chronological ontology, their time period directory, by extracting time words and dates from Library of Congress Subject Heading Authority Records. We extracted a simple word list from the Classification Schedules themselves, mostly from the schedule for history. To this, we added words associated with stylistic periods (such as Romanticism) and political periods (Latin Christendom). This simple list was sufficient for the collection because the time classifier depends largely upon dates and numbers (as does the metadata) and to only a minor extent on terms in the time ontology.

For subject indexing, WordNet is commonly used. But in light of WordNet's known drawbacks for information retrieval [Gabrilovich & Markovitch 2007], we looked elsewhere. The categories of the Library of Congress Classification System have been used to index bibliographic records [Larson, 1992] and for web pages [Wang et al., 2003] and [Prabowo et al., 2002]. Our work differed in that we used terms from the upper levels of the Library of Congress Classification (LCC) Schedules to supply terms for indexing maps.

How exactly did we compile the ontology for each browse category of the subject facet? The Library of Congress Classification, however, includes non-relevant words.[7] Rather than compile a simple list of terms as we did in the case of the time ontology, we used large portions of the Schedules and assigned weights to terms that were category indicators. We extracted phrases for subcategories of time and subject, where a phrase is defined as two or more words to be treated together rather than individually. Instances of phrases in the category "Medicine," for example, are first aid, intensive care, operating room, and physical therapy. Individually, the word "first" and the word "aid" mean something different than does the composite "first aid".

We gave phrases the highest weight because it has been determined that target documents that contain an exact query phrase are more relevant than documents containing merely the query words [Yeganova et al., 2009]. Phrases, as the very best indicators of classification category, were assigned 40 points, terms that were very good indicators of category were assigned 10 points, and good indicators of category were assigned 5 points. These values were arrived at and then adjusted through experimentation. Words that appear more than once in the metadata are weighted depending upon their number of occurrences.

Weights for words in the ontology

Phrases	40 x number of occurrences
Very good category indicator words	10 x number of occurrences
Good indicator words	5 x number of occurrences
All other words in the ontologoy	1 x number of occurrences

6 Evaluation

[7] For example, the category "Medicine" includes as a subcategory, "Pathology." The first entry for Pathology is "General works". Neither "general" nor "works" are medical terms, so rather than delete them, we simply do not weight them.

The procedure to verify the accuracy of an automatic classification algorithm is fairly standard. Overviews are provided by Keller [2001] or Oberhauser [2005]. A manually annotated corpus helps in evaluating the automated results. Evaluation might be on the basis of accuracy and error rate. Accuracy is a measure of whether all items that should be in a category have been retrieved into that category, and is also known as recall. To test our algorithm, we used a measure of simple accuracy. Other measures we might have used are precision, which considers whether those items retrieved for a category do in fact fit into that category, and the F measure, which comprehends both precision and recall.

The testing set consisted of 55 maps, collected just as were the training set maps from articles in diverse disciplines. All maps were unseen by the system. We created a benchmark by asking two people with professional indexing experience to classify each map according to the basic indexing rules used by the system. Further details of the evaluation are in Gelernter [2009]. The two people were not invariably consistent in their choice of categories, especially in the subject facet. Rather than choose between professionals' assessments, we declared *all* their classifications to be accurate, even when this allowed more than two categorizations per item.

The system classified the same items, and given the professionals' answers for scoring. All categories assigned by the system were considered right or wrong, except for those in subject, in which an answer could be partially right to compensate for category overlap. Statistics on system accuracy for each facet with respect to the manual benchmark are shown in the chart below (Fig. 2).

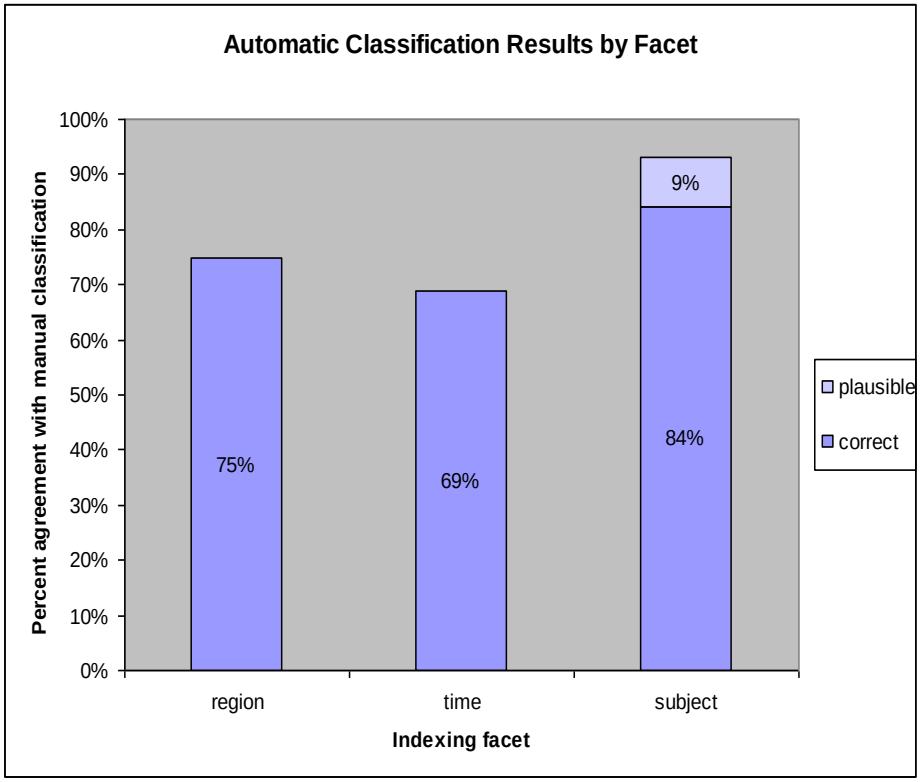

Figure 2. Evaluation of automatic region, time and subject classifiers in comparison to manual benchmark.

Our results run from 69%–84% (and even higher for the subject category, if you take into consideration that partially correct answers due to category overlap). We attribute this high accuracy to the fact that the main indexing fields of map caption, words-in-map, sentence that refers to the map, and the article title supply good indicators of region, time and subject category. Ontologies for region, time and subject that support the classifiers and augment the metadata with synonyms to improve matching are also a factor.

7 Conclusion and Future Work

The amount of data mined is key to develop these classification algorithms. Too little data will give too few terms to match with the domain ontologies, whereas too much data can introduce noise that will potentially match to the wrong domain ontologies. We learned that using the caption under the component graphic, some words in the graphic, the sentence in the article that refers to the graphic, and the article title provide words that are good category indicators.

In seeking a balance between too much or too little metadata, tests for region and time showed that it was better to prefer too little. This produced different situations per facet. To classify by region, there was never a lack of place indicator among the training set cases, so it seemed as though the quantity of metadata mined was adequate. To classify by time, when no specific date or era was mentioned in the mined metadata, it was assumed that the time was roughly the present, and so the publication date of the article could be used. To classify by subject, the mined metadata regions proved insufficient and it was necessary to search the entire article, essentially classifying the entire article rather than the map. Journal title might be additional information we could use for subject classification.

Another possible approach to improving classifications within each facet would be to consider correspondences between facets. For example, if a certain time period and subject tend to occur together, such as archaeology and pre-history, a rule might be established.

Ranking of retrieved maps, not a main focus here, would be a useful area for future work. The prototype already accommodates display sort by image size, crispness and color variety. Future work might consider the utility of ranking maps according to the number of scholarly citations to the map's host article, for example, or according to other map attributes. While it would certainly be useful to sort maps according to degree of map accuracy, it would be different to make such an inference. If cartography is produced by a particular mapping agency or publisher, we have some basis for judging accuracy by considering the source. In a map illustrating a journal article, however, it is hard to determine whether theme points were collected accurately, and if so, whether they were plotted accurately.

Extracting maps from articles could be a model for extracting other nuggets of interest. Automatic identification of the features to be extracted represents a line of research complimentary to this work on classifying extracted features.

Acknowledgements

Michael Lesk at Rutgers University coded the prototype, for which I am grateful. I thank also Heather Moulaison and Sarah Legins for the careful attention they gave to indexing the map sample which was used to evaluate the classification algorithms.

References

[Ahmed et al., 2009] *Ahmed, A., Xing, E.P., Cohen, W. W., Murphy, R. F. (2009). Structured correspondence topic models for mining captioned figures in biological literature. Knowledge Discovery in Databases KDD '09, June 28-July 1, 2009, Paris, France,* 39-48.

[Bishop, 1998] *Bishop, A.P. (1998). Digital libraries and knowledge disaggregation: the user of journal article components. Digital Libraries '98, Pittsburgh, PA, U.S.A.,* 29-39.

[Carberry et al., 2006] *Carberry, S., Elzer, S. Demir, S. (2006). Information graphics: an untapped resource for digital libraries. SIGIR '06, August 6-10, Seattle Washington, U.S.A.,* 581-588.

[Chen, 2008] *Chen, Y-L. (2008). A Robust Technique for Character String Extraction from Complex Document Images International Symposium on Information Technology, 2008. ITSim 2008, 26-28 August 2008, Kuala Lumpur, Malaysia, vol. 3,* 1-9.

[Dhar & Chanda, 2006] Dhar, D. B. & Chanda, B. (2006). Extraction and recognition of geographical features from paper maps. *International Journal of Document Analysis 8*(4), 232-245.

[Fan & Li, 2006] Fan, L., Li, B. (2006). A Hybrid Model of Image Retrieval Based on Ontology Technology and Probabilistic Ranking *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, December 18 - 22, 2006, Hong Kong, China,* 477-480.

[Futrelle & Nikolakis, 1995] Futrelle, R. P & Nikolakis, N. (1995). Efficient analysis of complex diagrams using constraint-based parsing. *International Conference on Document Analysis and Recognition (ICDAR-95), August 14-16, Montreal Canada*, 780-790.

[Gabrilovich & Markovitch, 2007] Gabrilovich, E. & Markovitch, S. (2007). Harnessing the expertise of 70,000 human editors: Knowledge-based feature generation for text categorization. *Journal of Machine Learning Research 8*, 2297-2345.

[Gelernter, 2009] Gelernter, J. (2009). Image indexing in article component databases. *Journal of the American Society for Information Science and Technology*, *60*(8), 1-12.

[Gelernter & Lesk, 2008] Gelernter, J. & Lesk. M. (2008). Is your map here? *The 17th International Research Symposium on Computer-based Cartography, September 8-11, 2008, Shepherdstown, West Virginia, U.S.A*, [13 pp.].

[Gelernter & Lesk, 2009] Gelernter, J. & Lesk, M. (2009). Text mining for indexing. *JCDL'09, June 15–19, 2009, Austin, Texas, U.S.A.*, 467.

[Guglielmo & Rowe, 1996] Guglielmo, E. J. & Rowe, N. C. (1996). Natural-language retrieval of images based on descriptive captions. *ACM Transactions on Information Systems 14*(3), 237-267.

[Hirano et al., 2007] Hirano, T, Okano, Y., Okada, Y. & Yoda, F. (2007). Text and Layout Information Extraction from Document Files of Various Formats Based on the Analysis of Page Description Language. *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Curitiba, Parana, Brazil, September 23-September 26*, Vol 1, 262-266.

[Jacobsen, 1968] Jacobsen, J. D. (1968). Interactive graphics in data processing: Geometric relationships for retrieval of geographic information. *IBM Systems Journal. Vol. 7*, issue 3-4, 331-341.

[Keller, 2001?] Keller, F. (2001?). Evaluation: Connectionist and statistical language processing. Retrieved June 4, 2008 from http://homepages.inf.ed.ac.uk/keller/teaching/internet/lecture_evaluation.pdf

[Kemp et al., 2007] Kemp, Z., Tan, L, & Whalley, J. (2007). Interoperability for geospatial analysis: A semantics and ontology-based approach. *Proceedings of the eighteenth conference on Australasian database, January 29- February 2, 2007, Ballarat, Victoria, Australia*, ACM International Conference Proceeding Series vol. 242, 83-92.

[Khan et al., 2004] Khan, L., McLeod, D., & Hovy, E. (2004). Retrieval effectiveness of an ontology-based model for information selection. *The VLDB Journal 13*, 71–85.

[Kim & Myaeng, 2004] Kim, P. & Myaeng, S. H. (2004). Usefulness of temporal information automatically extracted from news articles for topic tracking. *ACM Transactions on Asian Language Information Processing 3* (4), 227-242.

[Koike & Takagi, 2009] Koike, A. & Takagi, T. (2009). Classifying biomedical features using combination of bag of keypoints and bag of words. *International Conference on Complex, Intelligent and Software Intensive Systems CISIS 2009, 16-19 March 2009, Fukuoka, Japan,*. 848-853.

[Khurram et al., 2009] Khurram, K., Faure, C. and Vincent, N. (2009). Fusion of word spotting and spatial information for figure caption retrieval in historical document images. *10th International Conference on Document Analysis and Recognition (ICDAR '09), 26-29 July, Barcelona Spain*, 266-270.

[Larson, 1992] Larson, R. R. (1992). Experiments in automatic Library of Congress Classification. *Journal of the American Society for Information Science 43*(2), 130-148.

[Leidner, 2007] Leidner, J. L. (2007). Toponym resolution in text: Annotation, evaluation and applications of spatial grounding of place names. Unpublished doctoral dissertation, University of Edinburgh, United Kingdom. Retrieved January 8, 2008 from http://hdl.handle.net/1842/1849

[Li et al., 2005] Li, H, Zhao, T., & Li, S. (2005). Graphic retrieval based on limited semantics. *Proceeding of 2005 Natural Language Processing and Knowledge Engineering (NLP-KE' 05), October 30 –November 1, 2005, Wuhan, China*, 535-539.

[Liu et al., 2005] Liu, Y, Lazar, N. A., Rothfus, W. E., Dellaert, F. Moore, A., Schneider, J. & Kanade, T. (2005). Semantic-based biomedical image indexing and retrieval. Retrieved June 13, 2010 from http://www.irisa.fr/visages/demo/Neurobase/Didamic/d05.pdf.

[Lu et al., 2006] Lu, X., Mitra, P, Wang, J. Z. & Giles, C. L. (2006). Automatic categorization of figures in scientific documents. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, Chapel Hill, North Carolina, 129-138.

[Lu et al, 2007] Lu, X., Wang, J.Z., Mitra, P, & Giles, C. Lee (2007). Deriving knowledge from figures for digital libraries. *WWW 2007, May 8-12, Banff, Alberta, Canada*, 1229-1230.

[Mani et al., 2004] Mani, I, Pustejovsky, J., & Sundheim, B. (2004). Introduction to the special issue on temporal information processing. *ACM Transactions on Asian Language Information Processing 3* (1), 1-10.

[Martins et al., 2005] Martins, B., Silva, M. J., Andrade, L. (2005). Indexing and ranking in geo-IR systems. *GIR '05, November 4, 2005, Bremen Germany*, 31-34.

[Oberhauser, 2005] Oberhauser, O. (2005). *Automatisches Klassifizieren: Entwicklungsstand – Methodik – Anwendungsbereich*. Europäische Hochschulschriften. Series XLI Informatik, vol. 43. Frankfurt am Main: Peter Land.

[Pass et al., 2006] Pass, G., Zabih, R., & Miller, J. 1996. Comparing images using color coherence vectors. In *Proceedings of ACM Multimedia 96*. Boston, MA, U.S.A., 65–73.

[Perry et al., 2006] Perry, M., Hakimpour, F. & Sheth, A. (2006). Analyzing theme, space, and time: An ontology-based approach. *Proceedings of the 14th ACM International Symposium on Geographic Information Systems, ACM-GIS 2006, November 10-11, 2006, Arlington, Virginia, U.S.A.*, 147-154.

[Peterson, 2006] Peterson, M. P. (2006). Hypermedia maps and the internet. In E. Stefanakis, M. P. Peterson, C. Armenakis & V. Delis (Eds.) *Geographic Hypermedia: Concepts and Systems*. (pp. 121-136). Lecture Notes in Geoinformation and Cartography, Series Eds. W. Cartwright, G. Gartner, L. Meng & M. Peterson. Berlin: Springer.

[Petras et al., 2006] Petras, V., Larson, R. R., Buckland, M. (2006). Time period directories: A metadata infrastructure for placing events in temporal and geographic context. *Proceedings of the 6th ACM/IEEE CS Joint Conference on Digital Libraries, June 11-15, 2006, Chapel Hill, NC, U.S.A.*, 151-160.

[Prabowo et al., 2002] Prabowo, R., Jackson, M., Burden, P. & Knoell, H.-D. (2002). Ontology-based automatic classification for web pages: design, implementation and evaluation. *Proceedings of the 3rd International Conference on Web Information Systems Engineering, WISE '02, Singapore, December 12-14, 2002*, 182-191.

[Rowe, 2002] Rowe, N. C. (2002). Virtual multimedia libraries built from the web. *JCDL '02, July 13-17, Portland, Oregon, U.S.A.*, 158-159.

[Sanderson & Han, 2007] Sanderson, M. & Han, Y. (2007). Search words and geography. *Proceedings of the 4th ACM Workshop on Geographical Information Systems, November 9, 2007, Lisbon, Portugal*, 13-14.

[Sandusky & Tenopir, 2008] Sandusky, R. J. & Tenopir, C. (2008). Finding and using journal-article components: Impacts of disaggregation on teaching and research practice. *Journal of the American Society for Information Science and Technology 59*(6), 970–982.

[Syeda-Mahmood, 1996] Syeda-Mahmood T. (1996). Capturing shape similarity using a constrained non-rigid transform. In *Proceedings of the 13th International Conference on Pattern Recognition*. *ICPR '96 August 25-29, 1996 Vienna, Austria*, 617-621.

[Tan et al., 2009] Tan, Q., Mitra, P. & Giles, C. L. (2009). Effectively searching maps in web documents. *31st European Conference on Information Retrieval, April 6-9, 2009, Toulouse, France*. N.p.

[Tezuka et al., 2006] Tezuka, R., Kurashima, T. & Tanaka, K. (2006). Toward a tighter integration of web search with a geographic information system. *WWW2006, May 23-26, 2006, Edinburgh, Scotland*, 277-286.

[Verbert et al., 2005] Verbert, K., Jovanović, J., Gašević, D., & Duval, E. (2005). Repurposing learning object components. In R. Meersman et al (Eds.): *OTM Workshops 2005, LNCS 3762*, 1169-1178.

[Wang et al., 2003] Wang, Y., Hodges, J. & Tang, B. (2003). Classification of web documents using a naïve Bayes method. *Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence, Newark, New Jersey, U.S.A., 2-4 November 3-5, 2003*, 560-564.

[Yao et al., 2008] Yao, H., Etzkorn, L. H., & Virani, S. (2008). Automated classification and retrieval of reusable software components. *Journal of the American Society of Information Science and Technology 59(4)*, 613–627.

[Yeganova et al., 2009] Yeganova, L, Comeau, D. C., Kim, W., & Wilbur., W. J. (2009). How to interpret PubMed queries and why it matters. *Journal of the American Society for Information Science and Technology 60* (2), 264–274.

[Zhang et al., 2009] Zhang, J., Pan, H., and Yuan, Z. (2009). A novel spatial index for case based geographic retrieval. *Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human (ICIS 2009) November 24-26, 2009, Seoul, Korea*. ACM International Conference Proceeding Series; Vol. 403, 342-347.