Data Mining of Maps and their Automatic Region—Time—Theme Classification

Judith Gelernter
Carnegie Mellon University
Newell Simon Hall 4606
5000 Forbes Avenue
Pittsburgh, PA, USA 15213
+1 412.268.2805
gelern@cs.cmu.edu

ABSTRACT

The goal of this research is to organize maps mined from journal articles into categories for hierarchical browsing within region, time and theme facets. A 150-map training set collected manually was used to develop classifiers. Metadata pertinent to the maps were harvested and then run separately though knowledge sources and our classifiers for region, time and theme. Evaluation of the system based on a 54-map test set of unseen maps showed 69%–93% classification accuracy when compared with two human classifications for the same maps. Data mining and semantic analysis methods used here could support systems that index other types of article components such as diagrams or charts by region, time and theme.

Categories and Subject Descriptors

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

General Terms

Algorithms, Experimentation

Keywords

Geographic information retrieval, geospatial data, text mining, metadata harvesting, knowledge extraction, faceted classification, indexing, algorithms, classifiers.

1. INTRODUCTION

Finding maps has been hindered by arduous standards for cataloging spatial data and by unnuanced interfaces for digital map collections. Some digital map collections are organized by region, such that looking for a subject is confined to hit-and-miss typing in a keyword search box. In terms of map cataloging, collection level metadata is not gathered routinely, so there is no master index to point to a collection likely to contain a sought-after map. Catalog records for maps are likely to be incomplete because spatial metadata schemes tend to contain a very large number of fields which are time consuming to complete manually.

This research aims to solve one aspect of the problem of how to find digital maps [5]. The scope of the dissertation work is wide, encompassing metadata and map image collection, database collection, item classification, indexing, retrieval and interface design. While our research discusses creation of the system, this paper focuses on that aspect of the problem that concerns indexes of geospatial information and their relationship to the indexing of time and theme information. It describes how separate knowledge resources, and then separate classifiers were used for region, time and theme. The most specific examples come from the geospatial domain.

Below, §2 describes some related research on automatic indexing and how this research differs. §3 recounts how the database and classifiers were created and refined. §4 evaluates the classifiers, discussing results in §5 and concluding in §6.

2. RELATED WORK

2.1 Categories and ontologies

Categories could have been chosen based on how subjects could be grouped in a random sample, or they could have been chosen based on an external source, or they could have been generated automatically [10]. One aspect of this research examined how users ask for maps. The queries were then coded for region, time and theme [7]. Patterns in the coding seemed to indicate that region, time and theme should be indexing facets. The users' questions were not examined in detail to see how categories could be subdivided.

Knowledge resources for all three facets were adapted manually. Then each category was linked to the logically corresponding selection of knowledge resources in order to widen the scope of the queries and improve retrieval.

2.2 Database creation

Manually-generated metadata is accurate (if inconsistent among those who make it), but costly. One goal is to generate metadata automatically without sacrificing too much accuracy. A current approach to data mining for classification is found in [9]. Typically, the full text is combed for relevant metadata. Extracting geospatial information often has been a separate research problem [2]. In our research, specific areas of an article (such as the title) are identified as most likely to yield metadata that is relevant. We determined which fields would produce

The SIGSPATIAL Special, Volume 1, Number 1, March 2009

potentially the best region, time and theme descriptor metadata for a map [6].

The maps and map-metadata were mined manually given that it was the indexing of maps that was a dominant research problem. Databases of components extracted from articles have been recommended in user studies [21]. By creating a system that indexes component maps, we have taken a step toward this recommendation.

2.3 Indexing

Rules for each of the three classifiers in the dissertation were induced from manual indexing, with the exception of a few rules for geographic classification adapted from [1] and [14]. Similar to the information retrieval in the dissertation is the string matching algorithm described by [8], with an index consisting of triplets—words, classes and weights. Weights in the time and subject knowledge resources were assigned manually before any training documents were run or their metadata examined.

Assignment of items to categories, according to Sebastiani, might be single-label, or multi-label [22]. In single label classification, one item is assigned to one category. In multi-label, it may be assigned to more than one category. Our system assigns an item to more than one category only if the second highest score is within 25% of that of the top scoring category. We do not require a threshold value for category assignment.

Our survey of map-related queries showed that region, time and theme are parameters people ask for repeatedly [7]. The three-facet index has precedents such as [18]. So while the standard Geographic Information Retrieval (GIR) system indexes region and subject, we index time period in addition.

2.3.1 Region

Wang, Xie, Wang, Lu and Ma identified three major research directions in detecting geographic features in documents: 1) exploiting various geographic information sources 2) identifying and disambiguating place names (a problem from computational linguistics) and 3) developing effective computation approaches [23]. The present research does all three.

Different gazetteers have been used to improve retrieval. The Geographic Names Information System (GNIS) is used for an ontology in [14]. We used the World Gazetteer for a first pass, and Geonames as a second pass, both being freely available over the Internet in digital form.[1]

2.3.2 Time period

Chronological indexing here works the same way as geographical: features must first be identified, and then disambiguated. Temporal information extraction is discussed in [12]. No working ontology for time has been discovered. Petras, Larson and Buckland discuss the creation of a chronological ontology, what they call a time period directory, by extracting time words and dates from Library of Congress Subject Heading Authority Records [19]. They do not include algorithms for document indexing along with their directory.

2.3.3 Subject

A major difficulty in automatic classification by subject is that every metadata word becomes a possible discriminator, so as to create a very large feature space with many dimensions. Isolating frequently-occurring terms as potential discriminators is one way to winnow the feature space, and applying Natural Language Processing techniques to isolate nouns is another way to reduce features that has had not too promising results [11]. Our research, by contrast, limits the feature space by mining as metadata only that text believed to be relevant in classifying the maps. However, it was found in practice that the text of the entire article was potentially useful for theme metadata. Perhaps the name of the journal the article was found in would additionally be helpful in establishing dominant themes.

A commonly used controlled vocabulary for information retrieval is WordNet, and its drawbacks for retrieval are well known [4]. But there are alternatives. For example, the Library of Congress Classification System has been studied by [13] for classifying bibliographic records; and by [24] and [20] for web pages. But it seems as though the entire Library of Congress Classification system has not been used to supply ontology terms, as is done here.

2.4 Creation and evaluation of classifiers

Heuristics, in the general case, are made by examining a set of documents manually, spotting patterns, and writing these patterns into rules, or heuristics. Those heuristics are coded and the algorithm is run against the training set of documents. Then the algorithm is refined. Overly many corrections during the training phase risks overfitting the algorithm to that set. The procedure to verify the accuracy of a classification algorithm is fairly standard, and a good overview is provided by [17]. A separate set of test documents is used for evaluation, and the algorithm is run against these previously unseen documents. Algorithm performance has been measured on the basis of accuracy—or the percentage of items that have been retrieved for a category that fit that category, and error rate—the percentage of items in a category that do not fit.

3. METHOD

Creating the search engine, weighting the ontologies and indexing by subject have been detailed in [5]. This section explains how categories for user retrieval, data collection, and the indexing algorithms that link categories to corpus were made. Emphasis and examples here are drawn from indexing for region.

3.1 Categories and ontologies

Retrieval categories can be generated from the documents themselves, from users directly, by analysis of user queries, or from an independent taxonomy. An independent taxonomy was used here for reasons of continuing category stability and ease of collection expansion. Categories were created within each of the three facets of region, time and theme, but these categories were assigned for the top level only. The assumption was that category

[1] http://world-gazetteer.com/ and http://www.geonames.org/

subdivision would be left for future development, with each category heading requiring subdivision. Ontology terms have been divided according to categories. These terms improve retrieval to each category. While the time and theme ontologies will require more words as the database scales up, the double region ontology with its detail in reserve for the second pass is likely to be serviceable to a corpus of much greater scale.

3.2 Database

The 150-map training set consists of maps from articles in a wide range of disciplines. Most articles were in native .pdf format, with a few that were in .html, converted to .pdf. As mentioned above, the maps and metadata for each map were assembled by hand. A description of how the metadata were taken from particular fields (map caption, words-in-map, article title and sentence in the article that refers to the map) appears in [6]. Even though the work was done manually, two programs are being developed that should help automate it in future. Michael Lesk at Rutgers University is writing a program that will scan an article, and recognize and extract a map [3], [15]. Lee Giles at Penn State University is writing a program that will extract words from a map in a method similar to that in [25].

3.3 Classifier

Separate classifiers were devised for region, time and theme. Each classifier is composed of heuristics, that is, rules with no provable justification that have been found in most cases to solve a problem. Each classifier has its own domain ontology which expands the query and matches with the target metadata.

Common to region, time and theme classifiers:

1. Heuristics for selecting the bag of words
   Where should we take word from each article?
   How many words should we take?
2. Heuristics for how to filter words in the bag before matching
   Prefer certain words based on frequency of occurrence
   Avoid certain words based on spelling and capitalization
3. Heuristics for classification via string matching (using ontologies)
   Assign each item to up to two categories
4. Heuristics for ranking within a category for result listing
   Order results by semantic or geographical relevance, or
   Order results by date or file characteristics

Below is our region classifier as an example, with time and theme classifiers in [5]

/H1/ (Heuristic). Distinguish place words from non-place words

(a) Place words begin with capital letters
(b) Place words may be multi-word phrases in capital letters
(c) Place word(s) follow "Map of…"
(d) Place word(s) precede "map" or "eco-region" or "region" or "locale"
(e) The top 100 words from Geonames (such as bay, stream, center, hill, mountain, north, east, south, west) may indicate place, so that a word found within two words of one of these could be a geographical name.

/H2/ Location of metadata. Go through metadata regions searching for place words in the following order:

1. Primarily:
   map caption
   words in map (if any)
   article title
2. Sentence in article that refers to the map (if any)
3. Paragraph containing sentence that refers to the map
4. First sentence in article or abstract
5. First paragraph of article or abstract
6. Additional paragraphs

/H3/ Amount of metadata (for optimal recall and precision) Continue scanning metadata locations /H2/ from 1-6 until a classification has been assigned.

/H4/ Multivalent classification Match metadata in each location in /H2/with one or two classifications.

Example metadata: France, the American Colonies and the Revolutionary War. This item is classed both in Europe and in North America

/H5/ Preferences
In metadata

(a) Prefer place names that are repeated.
(b) Stem metadata as needed to match with place referent.
(c) If two places are found in one metadata location as listed in /H2/, prefer metadata not in parentheses
(d) Consider a name to be a place name if it is qualified within two words by another place name
    Example: In "Sydney, Australia," Sydney is not a person but a place in Australia
(e) If two places are in one metadata location as listed in /H2/, and one or both correspond to more than one place in the ontology/gazetteer, select the two places that have the shortest distance between them. This is called "geometric minimality" by Leidner [14].
    Example: Lincoln, Nebraska in metadata
    We assume Lincoln is a place and not a person because of (d)
    We assume Lincoln is in the United States and not in the U.K. because of (e)
(f) If a place named in the metadata could be resolved to either of two places in the referent, choose the place that is higher in the referent hierarchy
    Example: New Brunswick in metadata
    This will resolve to New Brunswick, Canada rather than New Brunswick, New Jersey.

/H6/ Removing noise in metadata

(a) Exclude phrases of the sort "published at/in [place]"
(b) Exclude newspaper names such as "X Times" or the "Y Chronicle" or the "Z Gazette"

/H7/ Classification

(a) Use table of country correlations between Geonames countries and system (MapSearch) categories
(b) Use table of waters for correlations between major oceans, seas and rivers and system (MapSearch) categories

(c) If no other place names are found but “world” “globe” or “global” appear in any of metadata locations in /H2/, assign to category “World”

(d) If three or more classification regions match, assign to category “World”.

Example metadata: In the map above, some of the goods move across the eastern Mediterranean and the others move across Anatolia, connecting western Europe to West Asia, East Africa, and India.
This item should be classified as “World”

/H8/ Ranking by relevance of scale
List first those matches that represent the smaller scale and correspond to the higher place in the gazetteer.

Example:
Browse query entered: North America and Europe
Retrieved: map of Manhattan and map of Spain
Ranking: map of Spain, map of Manhattan

/H9/ Ranking by relevance of data attributes
When relevance of scale is not clear from the metadata, when items are equivalent in scale, or when the user elects otherwise, other ways to rank retrieval results are offered. For example:

* Color – grayscale – black and white
* Most recent publication first
* Highest resolution first

3.4 System overview

The basic functions of most search engines are: crawling to locate data, mining data and metadata to assemble the database, the construction of an index to facilitate retrieval, processing queries and result ranking. To these functions, our system adds matching target metadata to an ontology that is able to improve retrieval and lend its hierarchy to semantic result ranking. A prototype of MapSearch is currently mounted on the web. [2]

4. DISCUSSION

Heuristics are at their most powerful when they are most abstract [16]. Our heuristics have proven robust on a test set one-third the size of the training set. How well the heuristics would scale to the many thousands of maps we would like the digital library to contain eventually, is difficult to say. Our map sample was extracted from journals on a wide range of topics. It is, however, not only the topics, but the text formatting and the writing style that will influence the “bag” words used for indexing, and this is a factor of journal type. As we widen the number of publications from which maps and metadata are mined, the heuristics will need to be retested and refined.

One way of fortifying heuristics would be to make correlations between facets. We have looked at correlations between theme and time facets for the 150-map training set.

Table 1. Correlation between map classifications for theme and time.

	Pre-history	Antiquity	Middle Ages	Early Modern	Modern
Agriculture & Food	0	0	0	0	9
Archaeology & Anthro.	9	2	1	0	1
Arts & Media	0	0	0	1	2
Commerce & Finance	0	0	1	7	9
History & Travel	0	0	9	14	9
Medicine	0	0	0	1	4
Military	0	0	0	0	10
Politics and Law	0	0	4	3	33
Religion & Education	0	0	4	2	6
Science	4	0	1	1	21
Society	1	0	0	2	17
Technology	0	0	0	1	12

The table lends evidence, for example, to support a heuristic that items classified in Technology and Transportation should be classified in the time period Modern, and that item classified in Military should also be classified as Modern.

Expansion of the collection will require also an expansion of category subdivision and of ontology granularity. Ontology expansion should improve results of classification of a larger collection just as will heuristics tuning. In an expanded database, maps from articles in proprietary databases will likely appear with only a link to the source article full text along with publisher instructions as to how to subscribe for access.

5. EVALUATION

An independent test set of comprising 54 maps was used to test the classifiers’ accuracy. Two participants were asked to construct the benchmark by manually indexing each item in the three categories. larger sample would have increased experiment validity, but the sample was limited for the sake of the participants who, as it turned out, required several hours and several rest breaks to complete the indexing.

Both participants had professional indexing experience. Each in turn was given a stack of articles with the maps flagged, and asked to index each map by region, time and theme. They were given a list of category labels and a brief instruction sheet

[2] http://scilsresx.rutgers.edu/~gelern/maps

explaining how to assign items to classes (one or two items per category, for example) as a guide, and a blank spreadsheet to record their categories.

Compilation of their results created a benchmark. The two were not wholly consistent in their choice of categories, especially in the subject facet. We adjusted for this by declaring all their classifications to be accurate. This left us with more than two categorizations per item, even when the rule for the system and the rule given for manual indexing were two categories per item at most.

The system was given the same test maps to classify, and its classifications were compared to those of the participants. All categories determined by the system were considered either right or wrong, except for those in theme in which partial credit was given to compensate for the overlap in category labels. An example of such an overlap is that Medicine is a Science, although the two are separate MapSearch categories. Comparison of the professionals' and the system's classifications is exhibited in the chart below.

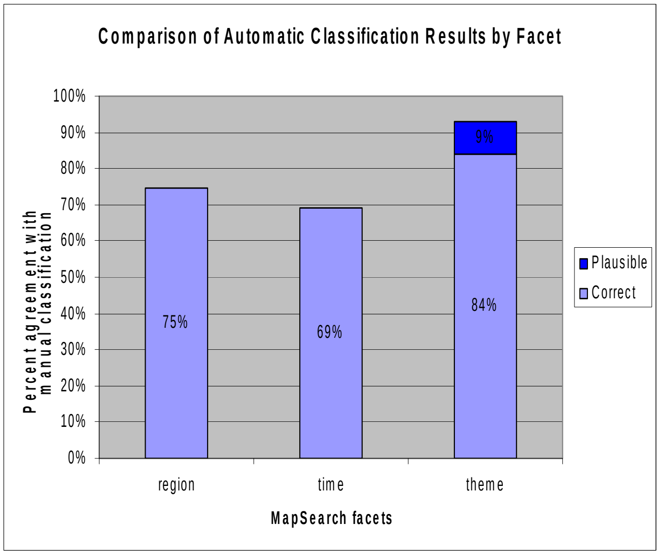

Figure 1 Automatic classification results by facet

6. CONCLUSION

In this research, maps for the corpus were extracted from journal articles. Classifiers for region, time and theme were created and knowledge resources were selected to aid in indexing each facet, with the region facet or geographic information retrieval aspect the focus of this paper. This research could be a model to aggregate diagrams, tables, drawings or other journal article components. Future research might include the accumulation of a larger map sample to further refine the algorithms, developing the automatic map mining program, expanding the time and theme ontologies and expanding subdivisions within each facet category.

7. ACKNOWLEDGMENTS

I am very grateful to Professor Michael Lesk of Rutgers University for advising me on the dissertation, of which this paper reflects a part, and for translating the algorithms into the programming language Perl. The Rutgers Academic Excellence Fund partially supported this research.

8. REFERENCES

[1] Amitay, E., Har'El, N., Sivan, R., & Soffer, A. 2004. Web-a-Where: Geotagging web content. Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Sheffield, United Kingdom, July 25 – 29, 2004), 273-280. DOI=http://doi.acm.org.proxy.libraries.rutgers.edu/10.1145/1008992.1009040

[2] Clough, P. 2005. Extracting metadata for spatially-aware information retrieval on the internet. Proceedings of the 2005 workshop on Geographic Information Retrieval, November 5, 2005, Bremen, Germany, 25-30 DOI=http://doi.acm.org.proxy.libraries.rutgers.edu/10.1145/1096985.1096992

[3] Entlich, R., Olsen, J., Garson, L., Lesk, M., Normore, L. and Weibel, S. 1997. Making a Digital Library: the contents of the CORE project ACM Trans. on Info Systems, vol. 15, 103-123 (April 1997) DOI=http://doi.acm.org.proxy.libraries.rutgers.edu/10.1145/248625.248627

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

[5] Gelernter, J. 2008. MapSearch: A protocol and prototype application to find maps. Doctoral Thesis. UMI Order Number: UMI Order No. pending, Rutgers University.

[6] Gelernter, J. and Lesk, M. 2008. Creating a searchable map library via data mining. Proceedings of the 2008 Conference on Digital Libraries (Pittsburgh, Pennsylvania, June 16 – 20, 2008). Joint Conference on Digital Libraries. ACM Press: New York, NY DOI=http://doi.acm.org.proxy.libraries.rutgers.edu/10.1145/1378889.1378997

[7] Gelernter, J. and Lesk, M. 2008. Is your map here? A well-designed interface will help you answer this question quickly. *The 17th International Research Symposium on Computer-based Cartography, September 8-11, 2008, Shepherdstown, West Virginia, USA.*

[8] Golub, K. 2006. The role of different thesauri terms and captions in automated subject classification. Proceedings of the ICCC/WIC/ACM International Conference on Web Intelligence, 961-965 DOI= 10.1109/WI.2006.169

[9] Graco, W., Semenova, T. and Dubossarsky, E. 2007. Toward knowledge-driven data mining. SIGKDD Workshop on Domain Driven Data Mining (San Jose, California, USA, August 12, 2007) 49-54. DOI =http://doi.acm.org.proxy.libraries.rutgers.edu/10.1145/1288552.1288559

[10] Jenkins, C. and Inman, D. 2000. Adaptive automatic classification on the web. Proceedings of the 11th International Workshop on Database and Expert System Applications (September 4-8, 2000), 504-511 DOI= 10.1109/DEXA.2000.875074

[11] Ke, H., and Shaoping, M. 2006. Text categorization based on concept indexing and principal component analysis. Proceedings of IEEE TENCON '02 51- TENCON '02. IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering Volume 1, 28-31 (Oct. 2002), 51 - 56 vol.1 no DOI.

[12] Kim, P. and 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 (December 2004), 227-242. DOI=http://doi.acm.org.proxy.libraries.rutgers.edu/10.1145/1039621.1039624

[13] Larson, R. R. 1992. Experiments in automatic Library of Congress Classification. Journal of the American Society for Information Science 43, 2 (1992), 130-148. DOI=10.1002/(SICI)1097-4571(199203)43:2<130::AID-ASI3>3.0.CO;2-S

[14] 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

[15] Lesk, M., Egan, D., Ketchum, D., Lochbaum, C. 1992. Better Things for Better Chemistry Through Multi-media. Proc. 8th Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research, Waterloo, Ont. 1992.

[16] Maxwell, C., Leaney, J., O'Neill, T. 2008. Utilising abstract matching to preserve the nature of heuristics in design optimization. 15th Annual IEEE International Conference and Workshop on the Engineering of Computer Based Systems, March 31 2008-April 4 2008, 287-296. DOI=10.1109/ECBS.2008.29

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

[18] Perry, M., Hakimpour, F. and Sheth, A. 2006. Analyzing theme, space, and time: An ontology-based approach. Proceedings of the 14th ACM International Symposium on Geographic Information Systems (Arlington, Virginia, November 10 - 11, 2006) ACM-GIS 2006, 147-154. DOI=http://doi.acm.org.proxy.libraries.rutgers.edu/10.1145/1183471.1183496

[19] Petras, V., Larson, R. R. and 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 (Chapel Hill, North Carolina, June 11 - 15, 2006), 151-160. DOI= http://doi.acm.org.proxy.libraries.rutgers.edu/10.1145/1141753.1141782

[20] Prabowo, R., Jackson, M., Burden, P. and 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, 182-191.

[21] Sandusky, R. J. and 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 (April 2008), 970-982. DOI=10.1002/asi.20804

[22] Sebastiani, F. 2002. Machine learning in automated text categorization. ACM Computing Surveys 34: 1–47. DOI= http://doi.acm.org.proxy.libraries.rutgers.edu/10.1145/505282.505283

[23] Wang, C., Xie, X., Wang, L., Lu, Y. and Ma, W-Y. 2005. Detecting geographic locations from web resources. In Proceedings of the 2005 Workshop on Geographic Information Retrieval (Bremen, Germany, November 4, 2005). GIR '05. ACM Press, New York, NY, 17-24. DOI= http://doi.acm.org.proxy.libraries.rutgers.edu/10.1145/1096985.1096991

[24] Wang, Y., Hodges, J. and Tang, B. 2003. Classification of web documents using a naïve bayes method. Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence (November 3 – 5, 2003), 560-564.

[25] Wu, V., Manmatha, R., Riseman, E. 1997. Finding Text in Images. Proceedings of the 2nd International Conference on Digital Libraries, July, Philadelphia, Pennsylvania, USA, 3-12. DOI= http://doi.acm.org.proxy.libraries.rutgers.edu/10.1145/263690.263766