Image Indexing in Article Component Databases

Judith Gelernter
Newell-Simon Hall #4606, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213.
E-mail: gelern@cs.cmu.edu

It is often necessary to compare data-rich charts, tables, diagrams, or drawings rather than the articles that contextualize that data. The objective of this research has been to create a database of non-textual components (here, maps) that are searchable independently of the articles from which they are taken, with the option to view the source articles. The method mines words from the articles that are near or associated with each component map, and these mined words become the basis of region, time, and subject indexing. The evaluation showed that automatic indexing of the component maps by these three facets works well, and indicates that a large-scale component database following this model is viable.

Introduction

New Type of Database

Some would profit from direct access to data-rich elements like charts or graphs that are contained *within* a document. This is shown in Sandusky and Tenopir's recent study of journal article components in which participants confirmed 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). Comparable databases in which the user searches for parts of a whole have been assembled for software components (Yao, Etzkorn, & Virani, 2008). Users look for sections of code that can be reused in another software. In our research, it is not code but map images that we index and classify, comparable to how biomedical images are indexed (Liu et al., 2005).

Significance of the Problem

Standard Machine Readable Catalog Records (MARC) for books in library catalogs contain a descriptive field that tells whether a book includes a map. The record does not describe the map, however. Articles rarely have even this level of indexing. This leaves thousands of information-rich graphics inaccessible outside of the articles in which they are embedded. This research posits a digital library of one type of component—a map. Access is essential. "One of the most important tasks in digital library management concerns the categorization of documents. Effectiveness in performing such a task represents the success factor in the retrieval process, in order to identify documents that are really interesting to the users." (Esposito et al., 2008, p. 127).

Related Map Digital Libraries and Access

MAGELLAN, described by Samet and Soffer (1998), is used for the acquisition, storage, and indexing of map images. Their earlier work, MARCO (Map Retrieval by Content), separated the maps into terrain and content layers, and indexed the maps by subject (Samet & Soffer, 1996).[1] DIGMAP holds images of historical maps and uses metadata from map libraries for indexing (Borbinha et al., 2007). Closer to our system is one described by Tan, Mitra, & Giles (2009) that includes maps extracted from documents, with metadata from the documents for indexing. Their system does not have ontology-supported indexes to aid retrieval for region, time, and subject, as does ours. And while their system indexes by field, it does not prioritize fields for indexing to improve relevance of results retrieved, as does our indexing procedure.

Scope and Outline

This article describes the making of a prototype database of article components, presently accessible at http://scilsresx.rutgers.edu/∼gelern/maps. We begin with background about the design of comparable systems and explain how ontologies potentially improve performance. Then we present our MapSearch system with a diagram and a description. We explain how text from an article associated with its component map that we use for metadata was

Received December 31, 2008; revised March 18, 2009; accepted April 3, 2009

 • Published online 18 June 2009 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.21107

[1]The theme of a map is also its subject, so theme and subject are used interchangeably in this article.

harvested. We describe how domain ontologies to improve retrieval were adapted, and how algorithms to automatically index article components were coded from hand-made heuristics. Indexing is according to region, time, and subject. Compiling indexes from region and time metadata is more straightforward than compiling indexes from subject metadata because each article has many subject terms and it takes an extra step to decide which subject term dominates, and hence, which subject category to use for classification. We include a detailed account of the weighting system we used for subject classification. Future research directions conclude the article.

Background

An article component collection, whether charts, diagrams, maps, or other illustrations, amounts to a database of images. Therefore, related work concerns classification and image retrieval. Two strains of machine learning that work for image retrieval are *content-based image retrieval* that analyzes low-level information extracted from features within the images (Nomiya & Uehara, 2007), and *semantic learning* that clusters images based on metadata, labels, or annotations (Liu et al., 2005). The difficulty with content-based analysis is what has been called a "semantic gap" between what the user asks for and the low-level features that the system extracts (Gosselin & Cord, 2004). In this respect, image indexing that uses metadata from journal article captions is more reliable.

Automatic Classification

Automatic classification is convenient when it would take too long to classify each item singly. Jain, Murty, and Flynn (1999) and Oberhauser (2005) review the literature on the automatic classification of items into categories. In preparation for creating a classifier algorithm, sample data are separated into a *training set* and a *test set*. Patterns distinguished in the training set are the basis of *heuristics*, or rules with some degree of generalizability for the *classifiers*. The classifiers group items into labeled categories. The match process between items and query may be mediated with linguistic sources such as dictionaries or ontologies. Analyzing misclassifications made during the training phase hones the algorithms, which improves classifier generalizability before previously unseen items in the test set are run for an evaluation.

The construction of an *inverted index* is fairly standard, and an overview of the process is described in the *Introduction to Information Retrieval* by Manning, Raghavan, and Schütze (2008, chap. 2). Our indexes are divided into fields for map caption, words in title, abstract, and so on. The words per field are in a random set, or a "bag." Each bag is tokenized (split into individual words), with stop words (mainly pronouns, articles, and prepositions) removed. Our documents were exclusively in English, so we were not concerned with accents or diacritics that would otherwise be normalized to ensure matching. We did, however, need to stem plurals.[2] Each time documents are added to the collection, a new index is created, in what is called dynamic indexing.

The popular approach to classifying documents automatically through the 1980s was knowledge engineering, or manually defining rules to assign items to categories (such as is done by a person indexing). By the 1990s, this lost favor to the machine-learning approach in which the rules were determined automatically from a set of preclassified documents. Types of classifiers include rule-based, probability-based, decision tree-based, multivariate regression-based, neural network-based, and nearest neighbor-based (Jain, Ginwala, & Aslandogan, 2004, p. 565).

Accuracy obtained from machine learning is comparable to that obtained by knowledge engineering, and does not require individual rule creation (Sebastiani, 2002, p. 2; Purpura & Hillard, 2006). Machine learning approaches require a very large document sample that we were unable to get, so the knowledge engineering approach was expedient.[3] It has been found that human interaction in the process is more effective when interspersed with the algorithmic process rather than when relegated to only before or after the automated parts (Nagy & Veeramachaneni, 2008). This human interaction has been categorized as *human-initiated* or *machine-initiated*, and described as *durable* (when used to alter system parameters) or *ephemeral* (when used to label patterns or modify results, for instance) (Nagy & Veeramachaneni, 2008, p. 237).

Why choose one method over the other? The drawback to pure knowledge engineering is its inflexibility (rules may require changing if categories are updated) and its lack of portability (rules may need to be reworked for each domain). In machine learning, on the other hand, the effort goes into the construction of the classifier builder (or learner) rather than into the classifier itself. But this may lead to inferior classifications.

Rules are derived from analysis of patterns within clusters as found in a training set of documents. Larger training sets should be more generalizable. Blanco-Vega, Hernández-Orallo, and Ramírez-Quintana (2005, p. 50) found, not surprisingly, that the size of the dataset, the number of rules that comprise the algorithm, and algorithm accuracy are related. Accuracy increases as well with training set size. While knowledge engineering methods tend to use a smaller

[2]Algorithms are available on the Web for stemming such as the Porter Stemmer. In MapSearch, only a few lines of code cover aspects of stemming. The suffixes –s, –ed, –ing, –ist, and –es are removed if its removal leaves a valid word, or if removing the suffix and adding an –e would leave a valid word. The same is true for adjectival endings –an, –ean, –ician, –ern, –ian, which are removed if a valid word remains. Future research will entail comparing stemming algorithms or adding more stemming code to what is in use for MapSearch to see whether better results may be obtained.

[3]We waited for some months for a large number of articles from JSTOR, but were ultimately disappointed.

training set with fewer rules, each rule potentially is more accurate because each is individually considered.

How does our work differ from most others? Ontologies as used here are not invariably used for automatic classification. In fact, the difference between methods is so pronounced that ontology-related indexing has been viewed as a stream of information retrieval separate from automatic indexing (Kabel, de Hoog, Wielinga & Anjewierden, 2004).

The Ontology and Its Role in Improving Retrieval Relevance

"Ontology" has been defined as a collection of concepts and their interrelationships that provide an abstract view of an application domain (Khan, McLeod, & Hovy, 2004, p. 71). Ontology models have different degrees of structure (Brusa, Caliusco, & Chiotti, 2006). The ontology may be able to improve retrieval over keyword search by using semantics, that is, word meaning, to expand matches with the words in the target metadata or words in the user query (Khan et al., 2004). The consequence is that results retrieved may be more relevant, and recall may increase.

Semantic relevance in the linguistics literature has been divided into two types: *relational similarity*, which amounts to a correspondence among parent–child relationships, and *attributional similarity*, which amounts to a correspondence among synonyms (Turney, 2006, p. 379). MapSearch uses both types of relevance to determine the retrieval set.

We use pre-existing hierarchically-arranged, controlled vocabularies as ontologies. In our choice of ontologies, in particular, our work differs from others. Rather than create an ontology (Khan et al., 2004), or use the general-purpose ontology WordNet (proven to be poor in information retrieval, as shown by Gabrilovich & Markovitch, 2007, p. 2329–2330), we adapted for subject indexing a controlled vocabulary from the library profession—Library of Congress Classification System. The World Gazetteer and GeoNames are our ontologies for region.

The ontology, transparent to the user, allows what may be called "smart" retrieval. "Smart" implies the ontology helps the system determine what the words mean rather relying on a possibly senseless word match between target and query term.

An Article Component Database

The MapSearch database consists of map images and metadata referring to those images, all of which have been extracted from journal articles and separated into three indexes for region, time, and subject. The entire process divides into pre-processing to identify the maps and metadata, and contextual processing of metadata to determine the classification. Procedures planned for the automatic mining of the data are described below, although for the purposes of this particular study, the maps were located and extracted manually.

Architecture of MapSearch

MapSearch will automatically index each component map by region, time, and theme, and it will implement smart (ontology-supported) retrieval by keyword term, browse category, or both keyword and category. Although MapSearch is still in development, its basic functions are diagrammed in Figure 1. Maps and the metadata associated with the maps are harvested from journal articles. When a user enters a browse category or keyword, the system compares the query to target terms in the metadata. Results are displayed according to semantic similarity by way of the ontologies, or according to other criteria such as map size or clarity.

Mining Maps

Discerning what is on the printed page has been called layout analysis. A review of the literature on physical layout analysis and segmentation methods that partition a page into text, image, and graphics is provided by Nagy (2000), while work on extraction and classification of diagrams in .pdf documents is found in Futrelle, Shao, Cieslik, and Grimes (2003). Layout analysis has been divided into physical layout analysis based on content presentation, and logical layout analysis based on content meaning (Marinai, 2008).

For the present research, maps were located manually, then clipped with Adobe Acrobat from articles (most articles are in .pdf and a few are in .html). Maps then were converted to .jpg. The whole process eventually can be automated with the help of a program that distinguishes a map from other kinds of graphics based on characteristics of the type of graphic. For example, the presence of straight lines will help distinguish maps and graphs. Such a program has been started by Michael Lesk at Rutgers, and independently by Lee Giles at Penn State. In the Lesk system, the page is divided into geometric zones to distinguish white lines between the text and other graphics. Tan et al. (2009) found that, in archaeological journals, most features larger than 1/3 of the entire page are maps. Error correction for this type of system would probably be manual.

Harvesting Metadata to Index the Maps

We use as metadata the map caption, words in the map (to be scanned separately), title of the article, and the sentence within the article that referred to that map. The choice of what metadata to harvest as dictated by preliminary experimentation is detailed in Gelernter (2008). We determined also the field order for reliability of metadata as indicative of classification category. Insight into priority of field relevance was carried into our classification heuristics.

A program has been created by Raymond Lu at Carnegie Mellon in continuation of this research. The program will separate title and caption, for example. To isolate words found in a map image, a layer extraction programs that will separate a map into base map and text has been started by Michael Lesk at Rutgers.

DOI: 10.1002/asi

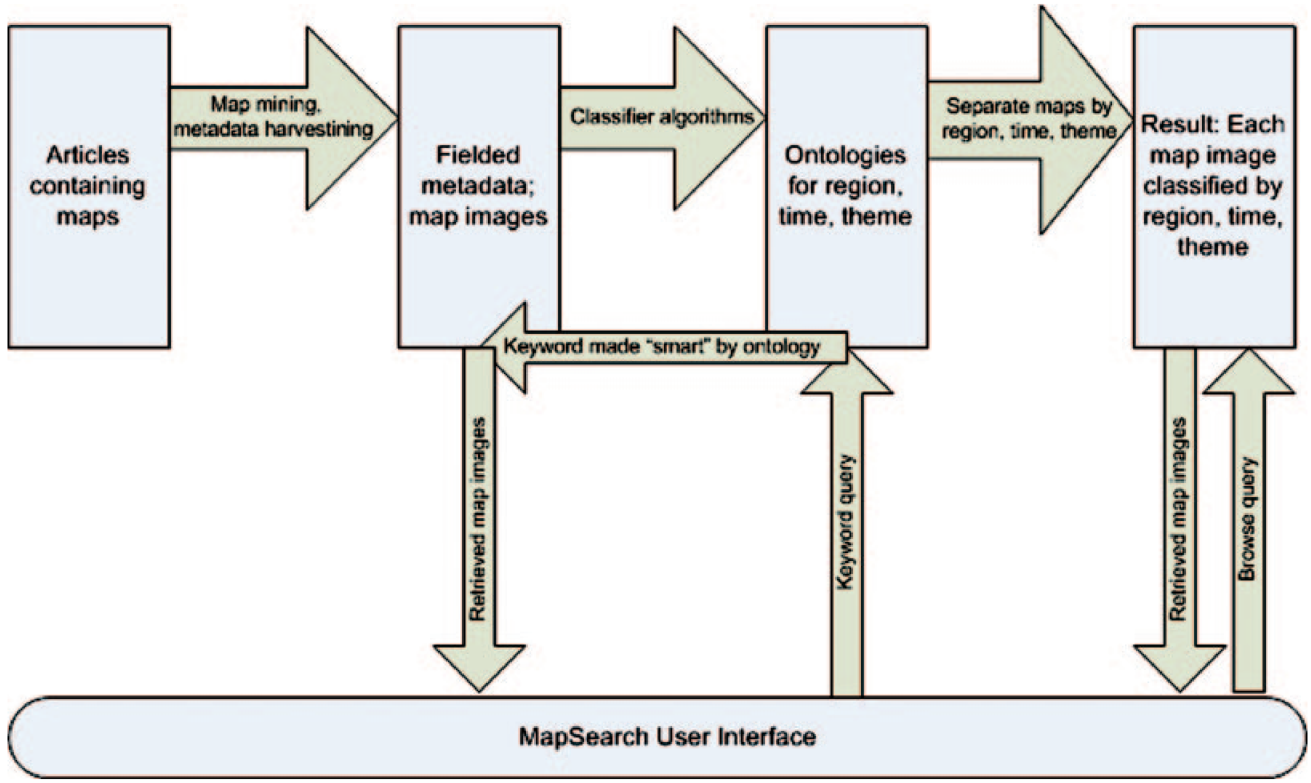

FIG. 1. Architecture of the MapSearch indexing and retrieval system.

How much metadata to harvest is yet another question. During trials with the training set, we check not only classifier accuracy (discussed separately below) but also the amount of metadata necessary for classification accuracy. Some measure of confirmation of how much metadata is needed is offered by Tan et al. (2009). They compare the retrieval value of metadata from the caption field, the referring sentence field, the caption and referring sentence combined, and the document page that contains the map. They found that the caption, referring sentence, and the combination of the two performed quite accurately, and that adding more metadata generally degrades performance.

Domain Indexes

It has been proposed that space, time, and theme should be considered as retrieval elements for a basic web search system (Perry, Hakimpour, & Sheth, 2006). Kemp, Tan, and Whalley (2007) call these three the "space–time–theme composite." If these three aspects could be used for referencing all types of information, as is argued by Hill (2006), they seem a solid foundation for indexing maps. Even so, examination of actual queries from people looking for maps reinforced the choice of triple-facet indexing (Gelernter, 2008).

MapSearch contains three browse facets, each supported by a separate index. One index contains metadata harvested per item with corresponding MapSearch categories for region, another index contains the same metadata with corresponding MapSearch categories for time, and the third index contains the same metadata with corresponding MapSearch categories for subject. Assignment of a given item to those categories requires a particular classifier and ontology for that domain.

Classification Categories

Different approaches are used in classification systems to determine classification category labels, such that the labels may precede or follow creation of categories. Preceding the creation of the categories is the use of labels from external controlled vocabularies, controlled vocabularies made internally to reflect a particular knowledge domain, or vocabularies derived from user terms as from user annotations (for example, Srinivasan, Pepe, & Rodriguez, 2009). Alternatively, labels might be made by hand to reflect the topic of machine-made clusters.

Our choice of category-label making for MapSearch reflected our limited quantity of data. We did not have a large enough data sample to cluster automatically, nor did we have access to a large sample of user annotations or other terms from which to create labels. For subject, we adopted controlled vocabulary for category labels. Controlled vocabulary external to the data has the possible advantage of being independent, so that the vocabulary terms will be valid for a much larger corpus. An in-depth description of how category labels were created for subject, time, and region facets is in Gelernter (2008). We limited the number of categories for each facet to between 10 and 15 because it has been shown

DOI: 10.1002/asi

that too many categories per level make browsing by category complicated visually (Theeramunkong & Lertnattee, 2002, p. 1).

How did we create category labels for subject? The Library of Congress classification which we used as a basis for subject labels was itself created for the fairly balanced collection of the Library of Congress (Leydesdorff & Rafols, 2009, p. 349). To make our labels, we started with the headings from the main schedules of the Library of Congress classification system, and condensed the 18 schedule headings to 12 logically related classes. The reason for condensing the labels was both to balance the data in our sample and to improve readability for a browse menu. Each of the categories will subdivide if we need finer-grained categories.

The beauty of the browse menu is that each category may unfurl separately if needed in an expanded system. As an example, the subdivisions for the Modern category of the time facet are active in the prototype. Subdivisions for the subject facet of the prototype are included to suggest what each category contains and to facilitate browsing. Would cooking be found within Agriculture and Food, for example? Selecting the browse category shows:

- Agriculture and Food
  Conservation and parks; Farming; Gardens; Hunting and fishing; Livestock; Veterinary medicine; Weeds and pollutants; Wildlife

As it turns out, cooking is a subdivision of "Society":

- Society
  Behavior: Of individuals, of groups; Lifestyle and family; Psychology and personality; Customs: Life cycle; Sports and games; Home and cookery; Hospitality; Language: Present and past languages; Linguistics

Ontologies for Each Indexing Domain

Different ontologies comprise different vocabularies, so the choice of ontology must affect retrieval. What did we choose, and why? We needed ontologies that we could download. We refined each domain ontology for comprehensiveness (extent of the vocabulary) and specificity (precision of the vocabulary) during iterative testing with the training set. Future research could include comparing classification results using different ontologies. But the high accuracy we have obtained suggests that the ontologies selected were good choices.

We refined the ontologies based on what we learned from misclassifications. For example, we began using the highly detailed GeoNames for region. We discovered it was misclassifying items due to matching metadata with local parks or obscure villages listed in GeoNames, because many of the same place names are found in different parts of the world. When we substituted the slimmer World Gazetteer and reserved the more detailed GeoNames for a second pass, our results improved dramatically. As another example, misclassifications by subject suggested that our ontology was not rich enough. Our attempts were unsuccessful to get a complete Library of Congress classification system in digital form beyond the classification outline available on the open Web.[4] So to add words to our subject ontology, we organized Machine Readable Catalog records from 130,000 randomly selected books according to the Library of Congress classification numbers to align them with our categories. We then removed their Library of Congress Subject Headings from the records and added them to the correct categories—an addition of about 800 words. We weighted phrases and individual terms as excellent, very good, and good indicators of category as before. The combination of classification label terms and subject headings classified many more training set items accurately.

Most indexing by time relied on dates mined from the articles. Because numerical data is unambiguous with respect to one of our time categories, semantic help was necessary rarely. This was fortunate in that no ready-made ontology for time was found. We compiled a short list of time words, extracted in part from the history schedule of the Library of Congress classification system but did not use it much for classification. For a more extensive corpus, the time ontology probably will need expanding as well.

Classifier Algorithms

Automatic classification is generally approached either by machine learning or knowledge engineering, as mentioned in the Introduction. The aim either way is to look at a quantity of data, discern patterns, and make rules based on those patterns. The best rules have the highest predictive accuracy for mapping independent variables (here, map images) to dependent variables (here, classification categories). We did not have the vast amount of training data necessary for machine learning, and besides, knowledge engineering is able to produce high rule-per-rule reliability.

We created three rule sets by examining the metadata for patterns that seemed to explain classification of the item by region, time, and subject. Heuristic rules for time and subject classification were created entirely by observation of the training set. In the case of region, observation was combined with principles reported by Leidner (2007).

Each of our classifiers works with its corresponding ontology. It is possible to classify from metadata to categories directly, without a controlled vocabulary or ontology. Even so, ontologies have been found to improve information retrieval (Kabel et al., 2004).

The three sets of heuristics were coded using Perl to make the three classifies. Coded heuristics may be applied sequentially in some order, or in parallel and all at once. An advantage of the sequential method is that rules can be ordered to place the rules of highest predictive value first, but a disadvantage is that lower-ranked rules are harder to interpret. While rule interpretation is easier with the parallel method,

[4] http://www.loc.gov/catdir/cpso/lcco retrieved January 2009.

DOI: 10.1002/asi

a higher priority rule may be undervalued in favor of a rule with lesser value in predicting a classification, possibly making the classification less accurate. We elected sequential ordering of heuristics primarily because we use metadata location to help judge reliability and boost classification accuracy.

We refined and debugged the classifiers by repeatedly running through the 150-item training set, examining misclassifications, and writing and then coding new rules in an attempt to improve classification accuracy. We continued to run the training data and alter the algorithms until we were fairly satisfied with the classifications that resulted. Then we halted this process and let the algorithms stand. At this point, a test set of 55 maps previously unseen by the system was run only once in an attempt to measure the accuracy of each classifier. The results are reported below in the Evaluation section.

Indexing by Subject

The indexing procedure adjusts to whether we index by region, time, or subject. Processing begins by removing metadata pertaining to the map and organizing it by field. The metadata are indexed three times: once matching to the region ontology that classifies into region categories, once matching to numbers that classify directly into time categories, and again matching to the subject ontology to classify into the subject categories. Subject indexing differs from region and time indexing because, among the metadata per item, a vast number of terms superficially could assign the item to very many categories. However, we do not want to classify each item into a large number of subject categories; instead, we need to reduce the dimensionality of the feature subspace. How to reduce the feature subspace by absolute means is a problem that occurs in many disciplines, and it is one that has been treated using statistical methods with some success (Dalalyan, Juditsky, & Spokoiny, 2008). Typically, for text processing, a system of weighting is used to prioritize some features over others.

Methods used to reduce the subspace include frequency of occurrence of the term in the document (Jain et al., 2004, for example), term co-occurrence (Liu, Wang, & Liu, 2004), term similarities as determined from the structure of the ontology (El Sayed, Hacid, & Zighed, 2007a), location of the term is found in the document (Macdonald & Ounis, 2006), or some combination of these.

Weighting to Determine the Subject

We prepared our subject ontology by assigning weights to individual terms. We decided that short phrases were the very best indicators of category. We added double stars to terms that were very good indicators of category, and single stars to terms that were good indicators. Ontology terms we did not single out in any way became a plain match with the target. Why did we score phrases so highly? It has been determined that target documents that contain an exact query phrase are more relevant than documents containing merely the query words (Yeganova, Comeau, Kim & Wilbur, 2009, p. 272). Words that appear more than once in the metadata are weighted depending upon the number of their occurrences. That repeated words are important is explained by Zipf's law, which predicts that a document's core vocabulary is likely to be repeated much more frequently than words less important to the core topic (Manning et al., 2008). Given Zipf's law, then, the more frequently-occurring words are more likely indicators of the dominant subject.

In binary indexing, each term weighs in at 1, despite its frequency, whereas in weighted indexing, a term assumes a different weight depending upon its importance such as its function or location in the document (Salton & McGill 1983).[5] Note that weights attached to ontology words make the classification results *appear* mathematically precise. We italicize appear because the ontology itself might be missing terms or branched unsuitably.

We used location of metadata word, frequency of word occurrence and category indicator potential of word (from the ontology) to determine the item's weight. We set our numerical weights by trial and error: if results showed that the items were being classified correctly, we kept our arbitrarily assigned weights. The subject ontology has words repeated among categories, as shown below (for instance, "experimental" appears in both Medicine and Science), so a match between just any ontology word and the target metadata scores minimal points. Determining an optimum numerical balance for weighting could be treated statistically, but many more examples would need to be considered than were available for this research.

In brief, the weighting system adds priority to terms in the metadata (by field) and to terms in the ontology (by strength of connection to a category). An item classified by subject without the benefit of a weighting system would appear in every category in which its metadata terms match ontology terms, while an item classified using a weighting system will be assigned only to the category of its dominant subject.

Example of MapSearch Weighting

Following is an example of how the weighting system assigns rules that prioritize metadata and determines which subject is dominant and should be used for classification. The MapSearch weighting system assigns point values to metadata words given each word's location in the article and the ontology term it matches with. Numerical values are assigned based on supposed relevancy to the map, so that fields closest to the map weight most heavily. For instance, a word with a weight of 8 might come from the caption whereas a word with a weight of 4 might come from the title; the "8" word is therefore a better indicator of classification for the

[5]We extracted some lines from our code, temporarily available at http://scilsresx.rutgers.edu/~lesk/t-class4.html, to demonstrate weighting. Enter any actual or imagined caption, map label, article title and sentence that refers to the caption and your item will be classified in to one of our MapSearch subject categories.

DOI: 10.1002/asi

map. Actual values were arrived at through experimentation. Below is an excerpt from our weighting system:

Weights for words in the ontology	
Phrases[6]	40 × number of occurrences
Word double star	10 × number of occurrences
Word single star	5 × number of occurrences
Word no star	1 × number of occurrences
Weights for words in the target metadata	
Caption words	8 × number of occurrences
Words-in-map	8 × number of occurrences
Title words	4 × number of occurrences
Referring sentence	2 × number of occurrences
Repeated words	1 × number of occurrences

The sample item to be classified is a map extracted from an article on the cultivation of sweet potatoes in early New Zealand.[7] It is a vector map in kilometer scale of New Zealand that uses diagonal hatch lines to indicate growing regions. Inset in this map to the right is a smaller map that shows New Zealand in its watery island context, just east of Australia. The inset middle left map in larger scale than the main map details the growing regions. Indications of the map theme are the key label "Kumara growing regions," and the "southern limit of Kumara gardening". These words are too small to be recognized by present scanning programs. Therefore, we rely on metadata in the article to determine the map theme automatically.

Metadata for classification come from text mined from the article: the map caption, article title, and sentence(s) in the article that refer to the map. The section of the output below lists the extracted metadata by field. The difference between what is found below and the text as it appears in the article is that punctuation, italics, diacritics, and capital letters have been removed to facilitate the match process.

CAPTION: fig 2 map of new zealand

TITLE: experimental archaeology gardens assessing the productivity of ancient maori cultivars of sweet potato ipomoea batatas in new zealand

REFERRING: of the four pre european kumara cultivars considered by maori informants to be of pre european origin or introduction yen 1963 33 rekamaroa was collected from ruatoria the east coast and the bay of plenty hutihuti from the east coast bay of plenty and northland taputini from northland and houhere from two locations in northland yen 1963 see fig 2

The next section of output shows the words from each metadata field that matched with words in the ontology, as organized by classification category. The categories are the browse categories in the MapSearch theme facet. The number after each word indicates its score. Notice in the Anthropology/Archaeology (ArchAnthro) category that "archaeology" scored 40, 4 points for being in the title field multiplied by 10 for its double-starred status in the Archaeology/Anthropology section of the ontology.

Division of category points by metadata field
CAPTION: fig 2 map of new zealand
History: zealand/8
Politics: zealand/8

TITLE: experimental archaeology gardens assessing the productivity of ancient maori cultivars of sweet potato ipomoea batatas in new zealand
Religion: ancient/4 archaeology/4
Society: ancient/4 garden/4
ArchAnthro: archaeology/40
Medicine: experimental/4
Science: experimental/4
Agriculture: garden/40 gardens/4
Commerce: productivity/4
Arts: potato/4
Politics: ancient/4 zealand/4
History: ancient/4 zealand/4

REFERRING: of the four pre european kumara cultivars considered by maori informants to be of pre european origin or introduction yen 1963 33 rekamaroa was collected from ruatoria the east coast and the bay of plenty hutihuti from the east coast bay of plenty and northland taputini from northland and houhere from two locations in northland yen 1963 see fig 2
Religion: collected/2 european/4 see/2
Society: origin/2
ArchAnthro: collected/2
Medicine: collected/2
Science: prey/4
Technology: coast/4
Arts: collected/2
Military: coast/4
Politics: see/2
History: coast/4 collected/2 european/4 see/2 two/2

The last section of the output as shown below tallies points per category. Again, the highest point value indicates the strongest belongingness to a category. Therefore, the dominant classification for this map is Agriculture as shown by its point value of 44.

Category	Totals
Agriculture	44
ArchAnthro	42
History	30
Politics	18
Religion	16
Society	10
Science	8
Medicine	6
Arts	6
Technology	4
Commerce	4
Military	4

Assignment: Agriculture and Archaeology/Anthropology

[6]Phrases consist of two or more words which the algorithm treats not individually but as a group. Instances of phrases in the category Medicine are intensive care, first aid, operating room, and physical therapy.

[7]Example from Burtenshaw, M. & Harris, G. (2007). Experimental archaeology gardens assessing the productivity of ancient Māori cultivars of sweet potato, Ipomoea batatas [L.] Lam. in New Zealand, Economic Botany 61 (3), 235–245.

DOI: 10.1002/asi

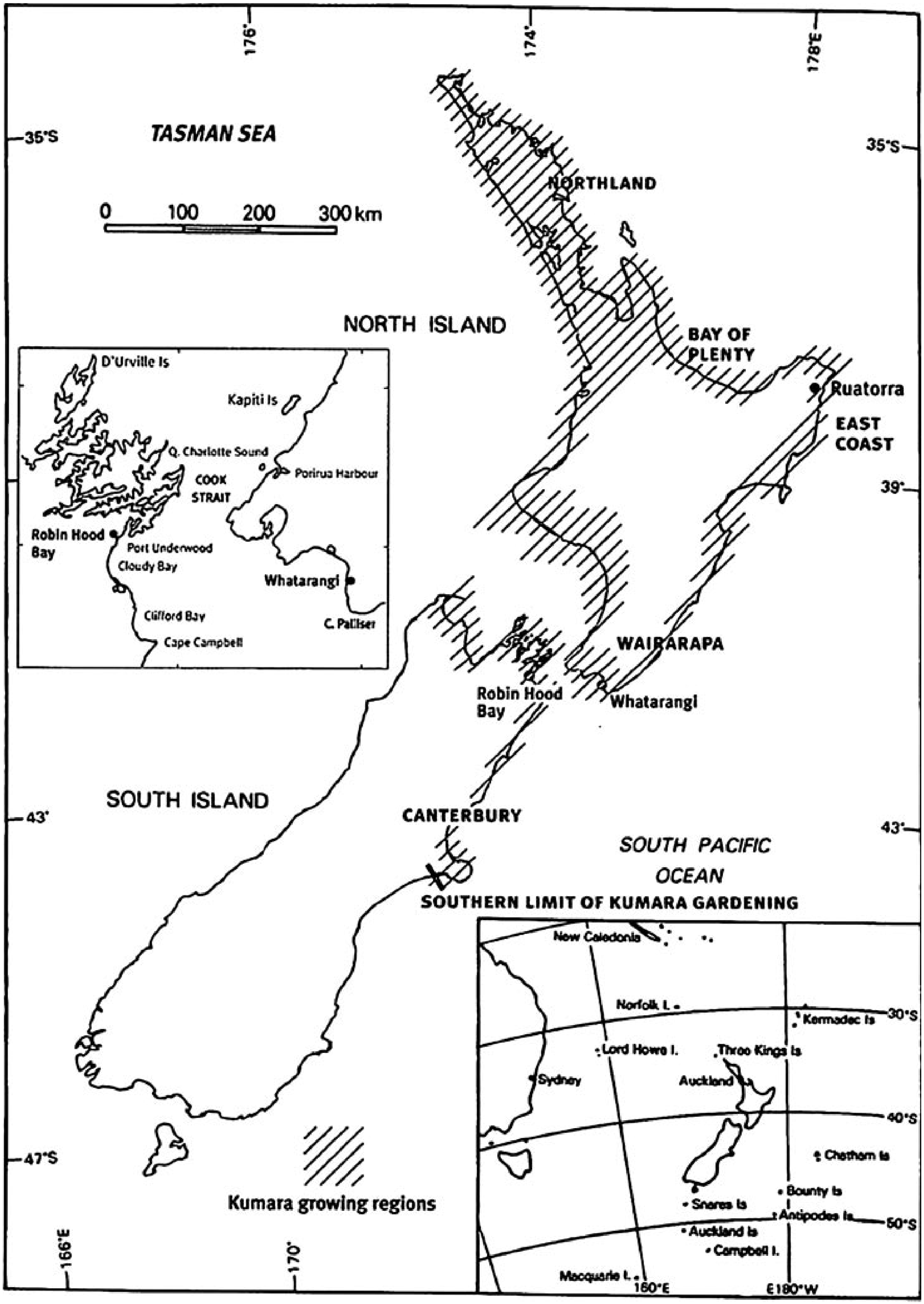

FIG. 2. Map of New Zealand.

The result is that this map was classified into two categories, Agriculture and also Archaeology/Anthropology. The section below discusses why and how we gave this map a second classification.

Multi-label Classification

Classification may be either single-label (also called absolute or hard) with each item classified into a single category, or multi-label (soft) with each item classified into one or more categories. Of single label and multi-label, the single label case is the more general of the two in that an algorithm for single label classification can be transformed into multi-label when the added categories are independent. Single label classification is more commonly discussed in the literature (Sebastiani, 2002).

Each category assignment in multi-label classification holds only an estimated probability of being a correct assignment. Absolute or hard classification generally is used in situations where the probabilities are not of primary interest (as for example, when the items are easily classifiable by humans), whereas multi-label is used to help evaluate or compare competing classifications (Wahba, 2002, p. 16524). Our research employs multi-label classification because many of the maps in the corpus can be classified into more than one subject category.

We decided, based on manual classification of the training set, that some maps require more than one subject category. We turned the multi-label requirement into a weighting system rule that the second highest scoring item must have scored within a certain percentage of the top scoring item. We determined that percentage by experimentation with the training set. Results showed that it is uncommon for the top two categories to score closely, say within 10% of one another, so that multi-label classification would rarely be indicated. On the other hand, it is quite common for the top two items to score within 50% of each other, which risks making items multi-label that rightly should belong to a single dominant category. We decided based on trials with the training set to adjust the percentage to 25% so that about a third of the items would be assigned multiple subjects. The actual percentage of maps assigned more than one subject is somewhat less than one third in that 59 of the 205 maps in the corpus have multiple subjects.

Sorting of Results

All retrieved results are relevant. But which among these is *most* relevant? It has been found that users are particularly likely to select the result listed first (Bar-Ilan, Keenoy, Levene & Yaari, 2009, p. 148. What is first in the display order is determined by the sort method. This section considers the relative ordering of results retrieved, acknowledging that "[c]omparing two objects relatively is still one of the biggest challenges and it now concerns a wide variety of areas in computer science. . ." (El Sayed et al., 2007b, p. 49).

Users may choose between sorting results according to semantic relevance to the keyword, or map size, image quality, color variety, or article publication date. Only the option of semantic relevance has a measure of subjective judgment, but it is also the option many prefer in looking for results. MapSearch determines semantic relevance by counting the number of steps between the query and the domain ontology. So that, for example, a query for Europe that retrieves maps of Paris, Europe, France in the result set will order the maps of Europe first, France second, and Paris third.

Absolute measurements for sorting are our options image quality, color variety, map size, and publication date. Image quality describes the sharpness of the map edges and is measured by converting each image to grayscale and measuring the between-pixel transition from black to white and white to black. Map size is measured by the total number of pixels. Color variety is measured by the number of different colors that appear in a map. Publication date refers not to date of the map itself, but instead to the date of the article that contains the map.

Evaluation

Methodology

We tested the effectiveness of the automatic classification by designing an experiment to compare manual and automatic classification of the same items. Human indexing of a test set of maps became the ground truth. The test set was composed of 55 maps assembled from journal articles covering a range of disciplines. A larger sample would increase experiment validity, but the sample was limited for the sake of the participants who were asked to index each item into three categories manually. Our planning was justified. As it turned out, the participants required several hours and several rest breaks to complete the work, and so a larger sample might have introduced indexer fatigue and possibly inconsistency, or even inaccuracy.

Two people with professional indexing experience classified the test items. Each participant was given a stack of articles with the maps flagged, a category list with explanations of what each category includes, a brief instruction sheet explaining how to assign items to categories (one or two categories per item), and a blank spreadsheet to record category assignments.

The items were ordered randomly. That random order was retained for both participants to help us keep track of the maps. Error could have been introduced by participant fatigue, making classifications at the end of the sample less accurate than at the beginning, or by participant inconsistency, with a person's conception of a category changing from the experiment's beginning to its end. Any negative consequences that might result from retaining the same order with two people were mitigated by the small sample size. We tried to keep the experiment short (with a relatively small number of items to classify), and encouraged participants to take breaks as needed. Also we gave participants freedom at the end of the study to return to the choices they had made at the beginning and make any changes they felt necessary.

DOI: 10.1002/asi

Procedure

When the participants had finished assigning our region, time, and theme categories to each of the 55 test items, their responses were compounded into a single list. The choice of classification was unambiguous in the majority of cases, and both participants selected the same category. When discrepancies did appear they were most often in the theme facet. The benchmark that was used for scoring thus in many cases contained more than two theme categories, but only one region or time category, and so the benchmark was more lenient in scoring for theme than for region or time. The outcome was that the system had wider latitude in choice of theme than in region or time categories—which probably explains why the agreement of results for automatic indexing of theme was highest of the three facets.

Results

The benchmark categories for the 55 test maps were entered into the system as the right answers. Then the system was asked to index the maps. System assignments were graded either right or wrong. The exception to this absolute grading system was that we gave partial credit to subject classifications in cases of predefined category overlaps.[8] The result was that MapSearch classified 75% correctly by region, 69% correctly by time, and 84% correctly by theme, with an additional 9% classified as plausible.

Discussion

The participants commented that, of the three facets, they had most difficulty assigning the theme subjects. Both relied on the article to assign a subject when the map seemed to be an illustration of the article, as suggested by the classification instructions. Another problem shared by the indexers was how to assign a time period to a modern map showing historical sites when the sites themselves fall into the categories Prehistory or Antiquity. Both participants elected to use the historical time period rather than the modern.

A larger training set would be the first step to improving classification of previously unseen terms. This is because analyzing more misclassifications would help us create new rules or alter the existing rules to improve generalizability so that similar items would be more likely to be classified correctly in the future.

How accurate does automatic classification need to be, ultimately? In other words, when should one stop tinkering in the hope of improving the algorithm? Larson (1992, p. 147) found in conducting an automatic classification experiment that only 46.6% of the sample could be classified correctly. Bates (1998, p. 1186) commented that it is typical for researchers to present a new system as 70% accurate. She pointed out that achieving the last 30% is vastly more difficult—all the more so as the collection grows.

Ideally, retrieval systems should be improved until they perform perfectly. Practically speaking, however, retrieval systems need only be reliable and perform useful work that could not be done otherwise. Take Google, for example. Google does not perform at 100% accuracy; in fact, its recall suffers when it misses results relevant to a given query. But even if an army of Web indexers could be found and paid, that army could not keep abreast of billions of new and changing web pages and return results instantaneously, as does Google. This excuses Google's imperfection. The same may be said of MapSearch. The concept of indexing article components is fairly novel. As the related work section shows, MapSearch has few functioning competitors. We conclude that imperfect results are better than none in retrieving maps from within articles.

Future Research

Automating Subtasks

Some of the data mining performed manually for the purposes of this research should be automated. Optical Character Recognition methods will be involved in the extraction of a map from its journal article, for example, and in the extraction of the words within a map to be used for indexing. Federated search protocols will be involved in adding new articles to the corpus.

Improving Tools and Mechanisms

The system is limited, among other factors, by decisions in the choice of classification categories, domain ontologies, and in the weighting of ontology terms. Each decision was made in light of preliminary testing, but we cannot test *every* alternative, so it is likely that potentially useful tweaks that might improve accuracy have not yet been discovered. It has been found that in machine learning, combining classifiers yields better classification, provided that the classification mistakes are somewhat independent (Manning et al., 2008). Performance of multiple classifiers has been attributed not only to independence but also to how classifiers are combined, with parallel (horizontal) combination used for high accuracy, and sequential (cascaded or vertical) mainly used for accelerating large category set classification (Liu & Fujisawa, 2008, p. 145). Combining our classifiers also might improve result accuracy, as discussed further in Gelernter (2008, p. 65).

Scalability

Some map-containing articles may not be able to be included in full text in a map retrieval database due to copyright restrictions. When full text is unavailable, the database

[8]The following interchanges were defined 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.

DOI: 10.1002/asi

could link to the journal publisher Web site. Users with access privileges would jump easily through the passwords to access the entire article, and those without would see publisher instructions as to how to subscribe.

There seems no reason to change the ontology domains, even for a scaled-up corpus. The two-gazetteer, two-pass system for region indexing is mature. However, for a large corpus, the time and subject ontologies will need supplementing in number and precision of words. To the time ontology could be added terms found in the geologic time scale (with eons, epochs, and stages), for example, and lists of style terms for describing periods of art and architecture (such as baroque). The subject ontology could be supplemented with words from within the schedules of the Library of Congress Classification.

The present browse categories in the prototype will continue to be applicable for a larger corpus, but the results will be less useful because many more relevant items will be retrieved per query. Therefore, finer-grained browse categories would help users view a manageable subset of what is available. An example of how this might work is presented in the subdivisions of the Modern time category in the prototype. The process of creating browse statistics for the users also requires automation. Presently, the intersection of various browse facets is precalculated so that, for example, choice of "Oceania" gives the user message that the system will retrieve nine maps, although coupling "Oceania" with the "Early Modern" period will net no maps. For maps added periodically to a much larger corpus, a formula will need to be devised to quickly calculate for the home screen this number of maps to be retrieved by the search.

MapSearch as a Model

We hope others will look to MapSearch for their research in retrieving article components within a single interface. Why aggregate article components? Those in agriculture need soil maps, water and crop maps, for example, those in finance profit from stock tables and past and present stock values charted over time, and those in the pharmaceutical industry want to see diagrams of chemical structures. Vertical search engines restrict search to a particular topic domain (Diligenti, Gori, & Maggini, 2002), and a subset of these are the Google custom search engines devoted to a field of study (such as economics) or a sphere of influence (such as the U.S. government).[9] Like these, a system that makes article components available would recommend itself to a group with common interests.

Conclusion

Some would benefit from direct rather than article-level access to quantities of graphical elements such as maps or charts or diagrams of chemical compounds. Our methods of metadata extraction by field (caption, words in the article title, etc.) and ontology-supported retrieval by region, time, and subject, proven successful in the MapSearch prototype, could translate to databases of other sorts of article components. Other researchers are encouraged to test and improve techniques presented here.

[9] http://www.google.com/coop/cse/examples/GooglePicks retrieved May 12, 2009

Acknowledgments

Thanks to Professor Michael Lesk, whose insight before, during, and after this project improved it immeasurably. Thanks also to the two anonymous reviewers. This research was supported in part by the Rutgers Academic Excellence Fund.

References

Bar-Ilan, J., Keenoy, K., Levene, M., & Yaari, E. (2009). Presentation bias is significant in determining user preference for search results—A user study. Journal of the American Society for Information Science and Technology, 60(1), 135–149.

Bates, M. (1998). Indexing and access for digital libraries and the internet: Human, database, and domain factors. Journal of the American Society for Information Science, 49(13), 1185–1205.

Beitzel, S.M., Jensen, E.C., Chowdhury, A., Frieder, O., & Grossman, D. (2007). Temporal analysis of a very large topically categorized web query log. Journal of the American Society for Information Science and Technology, 58(2), 166–178.

Blanco-Vega, R., Hernández-Orallo, J., & Ramírez-Quintana, M.J. (2005). Knowledge acquisition through machine learning: Minimising expert's effort. In Proceedings of the 4th International Conference on Machine Learning and Applications (ICMLA '05), (pp. 49–54). New York: ACM Press.

Borbinha, J., Pedrosa, G., Reis, D., Luzio, J., Martins, B., Gil, J., & Freire, N. (2007). DIGMAP—Discovering our past world with digitised maps. Research and advanced technology for digital libraries. Lecture Notes in Computer Science 4675 (pp. 563–566). Berlin: Springer.

Brusa, G., Caliusco, M.L., & Chiotti, O. (2006). A process for building a domain ontology: an experience in developing a government budgetary ontology. In M.A. Orgun & T. Meyer (Eds.), Proceedings of the Second Australasian Workshop on Advances in Ontologies (AOW '06) (pp. 7–15). Darlinghurst, Australia: Australian Computer Society, Inc.

Dalalyan, A.S., Juditsky, A., & Spokoiny, V. (2008). A new algorithm for estimating the effective dimension-reduction subspace. Journal of Machine Learning Research, 9, 1647–1678.

Diligenti, M., Gori, M., & Maggini, M. (2002). Web page scoring systems for horizontal and vertical search. In Proceedings of the 11th International Conference on the World Wide Web (pp. 508–516). New York: ACM Press.

El Sayed, A., Hacid, H., & Zighed, D.A. (2007a). Mining semantic distance between corpus terms. In I. Fundulaki et al. (Eds.), Proceedings of the ACM First Ph.D. Workshop in PIKM (pp. 49–54). New York: ACM Press.

El Sayed, A., Hacid, H., & Zighed, D.A. (2007b). A multisource context-dependent approach for semantic distance between concepts. In R. Wagner, N. Revell, & G. Pernul (Eds.), Database and Expert Systems Applications: Proceedings of the 18th International Conference on Database and Expert System Applications (DEXA) (pp. 54–63). Berlin, Germany: Springer.

Esposito, F., Ferilli, S., Basile, T.M.A., Di Mauro, N. (2008). Machine learning for digital document processing: from layout analysis to metadata extraction. In S. Marinai & H. Fujisawa (Eds.). Machine learning in document analysis and recognition (pp. 105–138). Berlin, Germany: Springer.

Futrelle, R.P., Shao, M. Cieslik, C., Grimes, A.E. (2003). Extraction, layout analysis and classification of diagrams in PDF documents. In Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR '03) (pp. 1007–1014). Washington, DC: IEEE Computer Society.

DOI: 10.1002/asi

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, J. (2008). MapSearch: A protocol and prototype application to find maps (Doctoral thesis, Rutgers University, 2008). Dissertation Abstracts International, 69-07A, 2495.

Gelernter, J., & Lesk, M. (2008). Creating a searchable map library via data mining. In Proceedings of the Eighth ACM/IEEE-CS Joint Conference on Digital Libraries (p. 452). New York: ACM Press.

Gosselin, P.H., & Cord, M. (2004). A comparison of active classification methods for content-based image retrieval. In L. Amsaleg, B.T. Jonsson, & V. Oria (Eds.), Proceedings of the First International Workshop on Computer Vision Meets Databases (CVDB '04), ACM International Conference Proceeding Series (Vol. 66, pp. 51–58). New York: ACM Press.

Hill, L.L. (2006). Georeferencing: The geographic associations of information. Cambridge, MA: MIT Press.

Jain, G., Ginwala, A., & Aslandogan, Y.A. (2004). An approach to text classification using dimensionality reduction and combination of classifiers. In Proceedings of the 2004 IEEE International Conference on Information Reuse and Integration (pp. 564–569). Los Alamitos, CA: IEEE.

Jain, A.K., Murty, M.N., & Flynn, P.J. (1999). Data clustering: A review. ACM Computing Surveys, 31(3), 264–323.

Kabel, S., de Hoog, R., Wielinga, B.J., & Anjewierden, A. (2004). The added value of task and ontology-based markup for information retrieval. Journal of the American Society for Information Science and Technology, 55(4), 348–362.

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

Kemp, Z., Tan, L., & Whalley, J. (2007). Interoperability for geospatial analysis: A semantics and ontology-based approach. In J. Bailey & A. Fekete (Eds.), Proceedings of the 18th Conference on Australasian Database (Vol. 63), ACM International Conference Proceeding Series (Vol. 242, pp. 83–92). Darlinghurst, Australia: Australian Computer Society, Inc.

Kirley, M., Abbass, H.A., & McKay, R.I. (2006). Diversity mechanisms in Pitt-style evoluationary classifier systems. E. Tiantaphyllou & G. Felici (Eds.). Data mining and knowledge discovery approaches based on rule induction techniques (pp. 433–457). Heidelberg, Germany: Springer.

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

Leidner, J.L. (2007). Toponym resolution in text: Annotation, evaluation and applications of spatial grounding of place names. Doctoral Thesis, University of Edinburgh, UK. Retrieved May 11, 2009 from http://hdl.handle.net/1842/1849

Leydesdorff, L., & Rafols, I. (2009). A global map of science based on the ISI subject categories. Journal of the American Society for Information Science and Technology, 60(2), 348–362.

Liu, C.-L., & Fujisawa, H. (2008). Classification and learning methods for character recognition: Advances and remaining problems. In S. Marinai and H. Fujisawa (Eds.). Machine learning in document analysis and recognition (pp. 139–161). Berlin, Germany: Springer.

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 April 21, 2009 from http://www.irisa.fr/visages/demo/Neurobase/Didamic/d05.pdf

Liu, Y.-C., Wang, X.-L., & Liu, B.-Q. (2004). A feature selection algorithm for document clustering based on word co-occurrence frequency. In Proceedings of the Third International Conference on Machine Learning and Cybernetics, Shanghai, 26–29 August 2004, Vol. 5, 2963–2968.

Macdonald, C., & Ounis, I. (2006). Combining fields in known-item email search. In S. Dumais, E. Efthimiadis, P. Hawking, & K. Järvelin (Eds.), Proceedings of the 29th Annual International ACM (SIGIR '06) Conference on Research and Development in Information Retrieval (pp. 675–676). New York: ACM Press.

Manning, C.D., Raghavan, P., & Schůtze, H. (2008). An Introduction to Information Retrieval [Draft]. Cambridge, UK: Cambridge University Press. Retrieved May 11, 2009 from http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf

Marinai, S. (2008). Introduction to document analysis and recognition. In S. Marinai & H. Fujisawa (Eds.). Machine learning in document analysis and recognition (pp. 1–20). Berlin, Germany: Springer.

Nagy, G. (2000). Twenty years of document image analysis in PAMI. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(1), 38–62.

Nagy, G., & Veeramachaneni, S. (2008). Adaptive and interactive approaches to document analysis. In S. Marinai and H. Fujisawa (Eds.). Machine learning in document analysis and recognition (pp. 221–257). Berlin, Germany: Springer.

Nomiya, H., & Uehara,K. (2007). Multistrategical image classification for image data mining. In K. Selçuk Candan & Z. Zhang (Chairs), Proceedings of the 8th International Workshop on Multimedia Data Mining: (associated with the ACM SIGKDD '07), International Conference on Knowledge Discovery and Data Mining (article no. 6). New York: ACM Press.

Oberhauser, O. (2005). Automatisches Klassifizieren: Entwicklungsstand – Methodik – Anwendungsbereich [Automatic classification: State of development-methodology-areas of implementation]. Europäische Hochschulschriften [European University Publications], Series XLI Informatik, vol. 43. Frankfurt, Germany: Peter Land.

Perry, M., Hakimpour, F., & Sheth, A. (2006). Analyzing theme, space, and time: An ontology-based approach. In S. Nittel & R.A. de By (Eds.), Proceedings of the 14th ACM International Symposium on Geographic Information Systems (ACM-GIS '06) (pp. 147–154). New York: ACM Press.

Purpura, S., & Hillard, D. (2006). Automated classification of Congressional legislation. In Proceedings of the 2006 International Conference on Digital Government Research, ACM International Conference Proceeding Series; Vol. 151 (pp. 219–225). New York: ACM Press.

Salton, G., & McGill, M.J. (1983). Introduction to modern information retrieval. New York: McGraw-Hill.

Samet, H., & Soffer, A. (1996). MARCO: Map Retrieval by Content. IEEE Transactions on Pattern Analysis and Machine Intelligence, 18(8), 783–798.

Samet, H., & Soffer, A. (1998). Magellan: Map acquisition of geographic labels by legend analysis. International Journal on Document Analysis and Recognition, 1(2), 89–101.

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.

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34(1), 1–47.

Srinivasan, R., Pepe, A., & Rodriguez, M.A. (2009). A clustering-based semi-automated technique to build cultural ontologies. Journal of the American Society for Information Science and Technology, 60(3), 608–620.

Tan, Q., Mitra, P., & Giles, C.L. (2009). Effectively searching maps in web documents. In M. Boughanern, C. Berrut, J. Mothe, & C. Soule-Dupuy (Eds.), Proceedings of the 31st European Conference on Information Retrieval: Advances in Information Retrieval. Lecture Notes in Computer Science, 5478, 162–176.

Theeramunkong, R., & Lertnattee, V. (2002). Multi-dimensional text classification. Proceedings of the 19th International Conference on Computational Linguistics, International Conference On Computational Linguistics (pp. 1–7). Morristown, NJ: Association for Computational Linguistics.

Turney, P.D. (2006). Similarity of semantic relations. Computational Linguistics, 32(3), 379–416.

Wahba, G. (2002). Soft and hard classification by reproducing kernel Hilbert space methods. Proceedings of the National Academy of Sciences of the United States of America, 99(26), 16524–16530.

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, 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.