P-127

CARTOGRAPHIC DATA-MINING

RUHSTORFER M., ECKRICH N.
Kartographie Huber, MüNCHEN, GERMANY

This article deals with cartographic data mining – especially the challenge of finding cartographic material in the Internet. Each map production starts with a research for the best available base material. This is the linchpin for quality in terms of positional and temporal accuracy as well as the richness of the content.

In pre-internet times it was usually sufficient to contact the responsible Land Surveying Office and license their data. Another possibility was to acquire data from the local city building authority or - for the task of producing touristic maps – to cooperate with a guidebook author. Today – in the age of Internet and Web 2.0 - there are several other possibilities. Nevertheless, the challenge of finding the best fitting data remains.

Main problems for this task are: Data discovery, Multilingualism, Spatial reference, the variety of formats

EXISTING SEARCH ENGINES AND THEIR USABILITY FOR CARTOGRAPHIC DATA MINING

The well-known search engines like Google, Bing or Yahoo! perform their indexing mainly based on different types of keyword-search. For images this approach would only work if all of them would be provided with a descriptive metadata set. This isn´t the case, therefore keyword indexing can neither provide information about image content nor spatial extent. Also Spatial Web Services, which – pushed through the implementation of INSPIRE - are increasingly important for the cartography branch, can´t be queried by ordinary search engines. Cartographic data mining will – and has to - be able to consider such services, and other "deep web" data sources as possible input.

IMPROVED MULTILINGUALISM WITH A SPECIALIZED THESAURUS

The latter of the problems can be easily addressed by the use of a thesaurus specialized on cartographic terms. A thesaurus consists of a collection of terms and their translation into different languages.

IMPROVED THEMATIC SEARCH CAPABILOTIES WITH FOCUSED CRAWLERS

Universal search engines have to deliver results for wide range of different topics. This has to be reflected in their ranking hits. Search engines which are optimized for a certain area of interest make use of so called Focused Crawlers. These software agents utilize content based analysis to rate the relevance of a website for a certain topic. Also the crawling is focused on URLs which are more likely relevant for the search queries without initial knowledge of the target website (based on URL- & Hyperlink-analysis).

Thereby Focused Crawlers are better suited to create high-quality and relevant document repositories for a certain domain compared to universal search engines, while saving computing power and disk usage.

The following search engines have been evaluated for their potentail in finding cartographic data

	Nutch	***Sphinx***	***Bingo***	***Combine***
Link-Topology	*Opic-Implementation*	*Yes*	*Hits, Page-Order*	---
URL-Order	*URL-Filter with rules*	*URL-Filter with rules*	*Link following for positive classification*	*URL-Filter with rules*
Additional Information		*SQL*	*Taxonomy positive/negative*	*Thesaurus Integration*
Classification	*Lucene Engine*	*Classifier*	*SVM*	*Automatic Classifier*
Protocols	*http, ftp*	*http*	*http*	*http, ftp*
Multilingualism	*Yes, Language identifier*	*No, Language identifier*	*Yes*	*No*
Open Source	*Apache Software License*	*Yes*	*Yes*	*GPL*
Platform	*Java, and other portings*	*PHP, Perl, Python, Ruby, Java etc.*	*Java VM*	*Linux*
Code	*Open Source*	*Open Source*	*Open Source*	*Open Source*
Development Activity	*High*	*High*	*Low*	*Low*
Documentation	*Good*	*Good*	*Good*	*Good*
File types	*HTML, TXT, PDF, and others*	*HTML, TXT,*	*HTML, TXT, DOC, PDF*	*HTML,TXT, PDF, DOC, PS, TeX*
Homepage	*http://nutch.apache.org/*	*http://www.sphinxsearch.com/*	*http://www.mpiinf.mpg.de/ departments/d5/ software/bingo/index.html*	*http://combine.it.lth.se*

The evaluation of the Focused Crawlers led to the decision for Kartographie Huber to rely on Nutch. Reasons are the active developer community, which hopefully will result in constant improvements, as well as the good documentation.

IMPROVED RESULTS WITH CONTENT-BASED IMAGE RETRIEVAL

The last issue which has to be addressed is the relatively limited image recognition capabilities of existing search engines.

• Text based methods: Text based retrieval methods determine the graphical primitives of images. These primitives are the distribution of colors within an image, the color histogram, the edge histogram, the dominant color, the image size, the color spectrum and the file format.

• Image based methods: So called QBVE-systems are based on an ancient approach of IBM. They make use of the fact that an example image will greatly improve the possibility to find similar images. This explains the acronym "query by visual example". It´s distinguished between systems which use example drawings of the user and systems which provide given example images.

• Combined Methods: There exist approaches which combine the afore mentioned methods. This eliminates many of their shortcomings. A functionality which isn´t provided till now – but should be implemented within the development of a cartographic data mining search engine – is the possibility to create a spatial reference of the image extent.

Within the research for cartographic data mining the systems shown in table 3 have been evaluated .

	Caliph&Emir	VizIR	IMGseek	WISE	Rummager	Brisc
Open Source Software	Yes	Yes	Yes	No	Yes	Yes
Multilingualism	No	No	No	No	No	No
Documentation	Good	None	Available	None	Available	Minimalistic
Colors	Yes	Yes	Yes	Yes	Yes	No
Structures	Yes	Yes	Yes	Yes	Yes	No
Textures	Yes	Yes	Yes	Yes	Yes	Yes
Metadata retrieval	Yes	Yes	Yes	Yes	Yes	No
Actuality	11-2010	6-2003	1-2009	10-2010	05-2010	Not anymore
Platform	Java	Java	C++, Python	-	Java	C#, .net 2.0

The evaluation showed that the Caliph and Emir software show promising results for their usability for our purpose.

RESULTS

The technologies described in this article will only improve the possibilities for cartographic data mining to a limited amount. It´s the combination of these technologies which we think make our approach interesting. A focused crawler will limit the results which have to be checked by the CBIR-System to a reasonable amount. Thereby the data repository will be "double-checked" to contain only the most relevant results. And the thesaurus for cartographic terms will enable the editor to retrieve information not only in languages he can master but also in additional languages relevant for the map production.

CONCLUSION AND FUTURE PLANS

Cartographic data mining is necessary to keep maps up-to-date and competitive to navigational data and other data providers. The classical maps have to be most precise to fit the markets needs and therefore cartographers need the best information available. To enable cartographers and editors to make best use of the internet they have to be provided with the right tools. In case of the Internet in our opinion this should be a search engine which combines focused crawling with Content-Based Image Retrieval functionality as well as multilingualism. Such a tool will help to restrict the information flood to a manageable amount. The problem isn´t that the information isn´t available. The problem is to know how to find it.