Work with Knowledge on the Internet – Local Search

Antonín Pavlíček, Josef Muknšnábl

Department of System Analysis, Faculty of Informatics and Statistics,
University of Economics, Prague, W. Churchill sq. 4, 130 67, Prague, Czech Republic
{antonin.pavlicek, muknsj}@vse.cz

Abstract. Authors are looking within their research grant new original web local search algorithm respecting specifics of Czech national environment. We would like to initiate further debate on topic. We are addressing three subtasks that include: identification of user geographical location, identification of web locality and final algorithm design working with these information altogether.

1 Preamble

A staggering pace of internet growth together with steadily increasing broadband penetration availability and general information literacy lead to more frequent internet usage. Such trend is not visible only in US but worldwide too - number of internet users and overall usage numbers constantly grow[1]. Although capabilities of engines and catalogs have improved significantly within last several years (especially after Google ranking algorithm arrival) they are still not perfect in terms of accuracy and relevancy. Typical areas where there is a potential for improvement are e.g. personalized and local search (in terms of geography and regions). Local search is a matter of an internal research grant that has been launched these days at University of Economics, Prague by us. A focus on that topic is not rare, especially in global scale, as several patents related to local search have already been filed[2] in US. Our main goal is to design and implement new web local search algorithm that will respect Czech national specifics and verify its function on local web page catalog Jihozapad.info. We would like to indicate possible ways of solution by the article in hope that some wider discussion bringing new ideas will be initiated.

2 Local Search and its possibilities

Web local search is a type of search when user is trying to find not only topic relevant but also locally (in terms of geographical distance) relevant web page/pages. Typically the users are searching for local/regional pages related to local businesses, local authorities or local events. Local search could be achieved by several ways. The most common one is by specification of country/state/area/district/city/village name (or other local information such as ZIP code) in query that is submitted to search site.

J. Pokorný, V. Snášel, K. Richta (Eds.): Dateso 2007, pp. 127–131, ISBN 80-7378-002-X.

Other one is that the search site recognizes user's physical location and will offer results relevant to recognized position only. The type of way is used depends on type of search site used. All major players in search engines/web catalog branch on global/Czech level offer local search tools. Let remind at least Google Maps, Yahoo! Local, MSN Live Search, from local Czech search sites mapy.cz and centrum.cz.

As latest numbers indicate an interest in local searching (geo-searching) is not a fiction or wish but a reality everyone has to count with. For example some recent poll, provided by comScore (Global Internet Information Provider) says[6] that more than 109 million of people performed about 849 millions of local searches in July 2006 which also represents 43% year over year increase. Most of the users, about 41 % were searching for items such as car rental office or dry cleaner [6]. A split by particular search engines / portals looks like this [6]: Google sites 29,5 %, Yahoo sites 29,2 %, Microsoft 12,3 %, Time Warner Network 7,1 %, Verizon Communications 6.6 %, YellowPages.com 3.9 %, Ask Network 2.7 %, Local.com 1.9 %, InfoSpace Network 1.9 %, DexOnline.com 1.4 %, all other sites 3.2 percent.

Such trend is confirmed by other polls and studies and is generally accepted and confirmed within whole IT/marketing industry[7].

3 Our initial conditions

As already mentioned within preamble we've decided to go the way of establishing new improved local search algorithm. That algorithm should be implemented in local web catalog/directory called jihozapad.info and its results verified within set of jihozapad.info registered www links.

Web catalog / directory jihozapad.info is primarily focused on area of South-West Bohemia (part of The Czech Republic). It contains primarily www links related to local subjects such as stores, companies, authorities etc. It geographically covers an area about 17 617 km² with about 1 180 541 inhabitants (population density is 67 inhabitants / km²). The catalog was launched in August 2005 and its 12 month-average unique visitor number is 1458 visitors per month. Catalog and its interface is primarily available in Czech, other available language is German, as covered area directly neighbor with Germany. Surprisingly most of visitors is from US (62 %)[1] followed by The Czech Republic (16%). Number of German visitors is quite insignificant (about 1%)!

There are 121 registered users and 1148 registered local web links. Locality is in catalog presented by possibility to determine district within which searching should be performed (there are actually four main districts called Klatovy [KT], Domažlice [DO], Strakonice [ST] and Plzeň-jih [PJ]).[2] A catalog has no true search engine at the moment, all links are added and registered only by registered users (approved by portal administrator)

[1] Robots are excluded.

[2] Information about district is available for all registered links. It is a mandatory attribute.

4 Problem decomposition

4.1 Identification of user geographical location

is also called geo-location. Typically geo-location of users is derived from their IP addresses (or MAC address). Such service is often available on commercial basis (such as IP2Location, MaxMind etc.). However it will not be very likely our case due to from our perspective high cost of such services. We'll try to discuss that with local providers and agree on some cooperation at this point. A level of details that we can obtain from IP address will depend on quality of service/database it will be for such purpose used. The easiest way task is to obtain name of the country (IP registrars supply that information for free), the more difficult is to get some other details such as region state, province/district, city, latitude/longitude etc. Other possible and used way of determining locations of user is to use information that user provided us during portal registration (such as address, ZIP code, phone numbers, GPS coordinates etc.). The problem is that number of registered users will be very likely much smaller than number of visitors, so its capabilities will be rather limited comparing the first mentioned method. Very likely combined approach will be chosen.

4.2 Web page link and its relation to particular geographical area

There are many ways that can help us to determine web page locality. We've thought about following, so far:

Use information provided by web page owners: there is information about district for each registered link right now in Jihozapad.info. We do not consider this as fully sufficient and there has been implemented an improvement leading to make location of www links more precise these days. We still will come from information that will be entered by user during link registration but this information will be more detail and will be expressed in a standard way. As an appropriate standard we have chosen split into geographical areas based on EU legal framework for the geographical division of the territory of the European Union also know as NUTS [8]. There will be a possibility to enter for one www link more geographical locations as one www link may represent a company with different stores within region (for example www.welstam.cz). Following NUTS information will be gathered:

- NUTS1_uzemi: Česká republika (same for all registered link)
- NUTS1_kod: CZ01 (will be the same for all registered link)
- NUTS2_oblast: Jihozápad (will be the same for all registered link)
- NUTS2_kod: CZ03 (will be the same for all registered link)
- NUTS3_kraj: Jihočeský kraj / Plzeňský kraj
- NUTS3_kod: CZ031 / CZ032
- NUTS4_okres: Strakonice / Domažlice / Klatovy / Plzeň-jih
- NUTS4_kod: CZ0316/ CZ0321 / CZ0322 / CZ0324

Such information will be also enhanced by particular address in form: City/Town, Street house number, ZIP code. Also information about latitude/ longitude and altitude will be gathered include precise GPS coordination (WGS-84). We strongly hope that all gathered information will help in providing better result on local search.

Use local specialties from web page content: Such approach is applicable in the case of automated geo-spatial search engine (which is apparently not the case of such improvement because of time restrictions). The idea is to search particular web page (include all subpages) for existence of unique local words such as addresses parts (village/town/district/area names), dialect words, ZIP codes, dial codes and derive web page locality from occurrence frequency of such words (or via other algorithm). Situation in that might be complicated by fact that many addresses can be found on webpage. However as Jihozapad.info is strictly oriented on region of South-West Bohemia (districts Klatovy, Domažlice, Strakonice, Plzeň - Jih), found addresses from other regions could be ignored. Similar algorithm to "Geographic Scope[3]" developed by Kyoto researches could be applied or other algorithms coming of data-mining techniques such as association analysis, clustering methods[4] etc.

Cooperation with local webhosting providers: identification and focus on local webhosting servers where there can local content will be very likely stored. For example local Webhosting provider ŠumavaNet contains lot of regionally oriented web pages. Webhosters also could become partners in gathering locally oriented content, via some unified interface for example.

Supporting and propagating standards helping in geo-location: jihozapad.info should be prepared to extract web page locality from some HTML-GEO formats/protocols such as Microformats hCard [5] (extension of item a) or cooperate in exchange of geo-spatial data associated to GIS systems distributed in a set of predefined formats. It would significantly improve catalog accuracy however because of timing restrictions it will not be possible the case.

Although there are many ways by which we can determine web page locality, no one of them guarantees for 100% the result. The reasons for that may vary. Many regional web pages, even those locally oriented don’t contain any significant information about their origin (they can be just topic oriented). Many of them are locality independent and finally quality of locality information derived by using methods mentioned above doesn’t need to be sufficient for locality determination.

4.3 A final search algorithm structure

These days jihozapad.info offers its users „district“ level of detail in relation to registered web pages. This granularity is of course not sufficient for being real locally oriented search site and improvements have already started to be implemented. Having all information about users accessing jihozapad.info and locality of registered web pages we can think about appropriate algorithm. At this moment we think of some kind of Google style ranking algorithm with different weights for particular levels of granularity (region/district/town/street) and specific metrics for deriving web page importance in given area (pages with links from other pages same district/region/town etc. would be considered as more relevant).

5 Conclusion

To find a good algorithm for local searching is a complex task that combines methods from many areas such as data mining, web pages constructions, search engine principles etc. We are tat the beginning right now, all methods mentioned in our article would help us, finding and optimal balance that will provide the most relevant and accurate result will be matter of real algorithm tuning on real data.

References

1. Market Research, *Internet World Stats – Usage and Population Statistics* [online]. [cit. 2006-12-20]. URL: <http://www.internetworldstats.com/stats.htm>.
2. SLAWSKI, William, *Assigning Geographic Locations to Web Pages* [online]. [cit. 2006-12-28]. URL: <http://www.seobythesea.com/?p=386>
3. YAMADA, Naoharu – LEE, Ryong – KAMBAYASHI, Yahiko. *Classification of Web Pages with Geographic Scope and Level of Details for Mobile Cache Management, Proceedings of the Third International Conference on Web Information Systems Engineering (Workshops)* 0-7695-1754-3/02, 2002 IEEE, [online]. [cit. 2006-12-20]. URL: <http://csdl.computer.org/dl/proceedings/wisew/2002/1813/00/18130022.pdf>
4. HAN, Jiawei – KAMBER, Micheline. Data Mining: Concepts and Techniques. San Diego,(CA), USA: Academic Press, 2001, 550 s., ISBN 1-55860-489-8
5. *hCard Description* [online]. [cit. 2006-12-20]. URL: <http://microformats.org/wiki/hcard>.
6. *comScore: Local Web Searching Soars* [online]. [cit. 2006-10-02]. URL: <http://www.mediaweek.com/mw/search/article_display.jsp?vnu_content_id=1003188359&schema=>.
7. *New Developments in Local Search, Part 4* [online]. [cit. 2003-11-19]. URL: <http://www.clickz.com/showPage.html?page=clickz_print&id=3110641>.
8. *Common classification of territorial units for statistical purposes* [online]. [cit. 2006-02-06]. URL: < http://europa.eu/scadplus/leg/en/lvb/g24218.htm>.