Adding Geographic Scopes to Web Resources

Mário J. Silva Bruno Martins Marcirio Chaves Nuno Cardoso
Ana Paula Afonso
Faculdade de Ciências da Universidade de Lisboa
1749-016 Lisboa, Portugal

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

Geographic IR, Web Mining, Geographic Entity Recognition

1. INTRODUCTION

In the past decade, Web search engines evolved from using classic IR models to infering relevance from the analysis of the Web graph [5]. We propose to further improve the quality of Web searches, by integrating the semantic knowledge that can be infered for Web resources.

In this paper, we describe our research on the identification of geographic scopes for Web pages. We define the geographic scope of a page as the region, if it exists, where more people than average would find that page relevant. Once scopes are assigned, searches may specify, explicitly or implicitly through context information, that the pages of interest must have a given geographic scope. Or, that the more relevant results are those "nearby" the searcher's geographic location.

We are adding support for this kind of searches to the next version of tumba!, a fully-functional search engine, which has been operating as a public service since 2002 [25]. Tumba! (`www.tumba.pt`) indexes the sites of interest to the people related to Portugal [9].

Statistics collected in our system over the past two years confirm previous findings that there is good potential for using geographic information on the Web:

- Geographic information is pervasive on the Web. A study over 3,775,611 documents found 8,147,120 references to the 308 Portuguese municipalities, an average of 2.17 per document [19].
- Geographic entities are frequent in user queries. A study of the search engine logs found that approximately 4% of the queries contained the names of the same 308 municipalities [6]. If we considered names of localities, landmarks or streets, this percentage would increase. It would certainly also increase if our engine considered geographic semantics and proximity when giving results.

Our approach incorporates concepts defined within the framework of ongoing work on the Semantic Web, such as Dublin Core, RDF and Topic Maps [1, 3]. The harvested Web data is analyzed with linguistic and statistical variants of natural language processing [16], data mining [15], Web graph analysis [7], and information extraction [10]. As 73% of the pages maintained in our engine are written in Portuguese [19], we plan on using Portuguese-specific NLP techniques for extracting the geographic information present on Web resources.

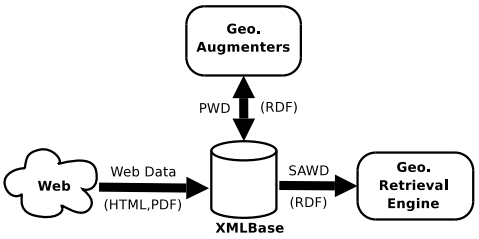

Figure 1: Overview of the Project.

The global view of our framework for assigning geographic scopes to web pages is shown in Figure 1. Web data is harvested into XMLBase, our Web data management system [2]. XMLBase offers the capability to process large document collections in parallel. Its components include a crawler, separate data and meta-data repositories, and an XML data access manager that integrates with XML query engines [14].

While "crawling" Web documents (HTML, PDF, etc.), a parser digests their content into RDF representations, which are then stored in the repository. We call these *Purified Web Documents* (PWD), to express that Web data, before becoming available for analysis, is cleaned into a collection of well-formed XML documents, organized under a common schema. This is an important first step, as handling Web data usually involves processing badly formatted information, with markup errors introduced by hand-editing documents or buggy authoring tools. The resources contained in a PWD include: i) meta-data properties extracted from the documents; ii) text tokens, sentences and HTML structural markup; iii) hyperlinks to other pages.

Purified Web Documents are the starting point of a chain of transformations that tag the named entities present in these doc-

Workshop on Geographic Information Retrieval, SIGIR '04, Sheffield, UK

uments, and then incorporate other knowledge to eventually assign the scope to the initial PWDs. Each of these transformers, called *augmenters* [11], can be thought as a domain-specific expert. The (geographical) knowledge is embedded within the documents as additional RDF resources, and we call these enriched resources *Scope Augmented Web Documents* (SAWD).

Finally, the indexing and retrieval components of the search engine will not only match text tokens from the Web resources, but also "geographical scope" information generated by the augmenters.

2. IDENTIFICATION OF "SCOPES"

In a typical geographic scopes assignment process, an augmenter might, for instance, look at the PWD and extract the geographical entities mentioned therein, producing a "geographically tagged" extension. The next augmenter might infer that pages mentioning the phrase "Portuguese capital" are in fact mentioning the geographic entity "Lisbon". A third augmenter might then propagate the extracted geographic entities through the linkage information among the pages, and so on.

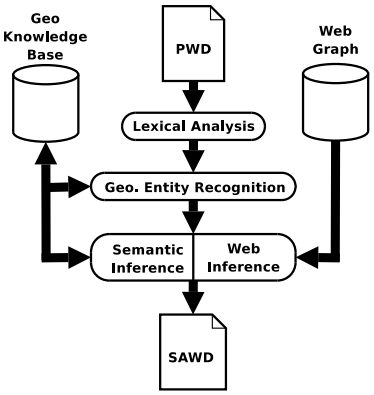

Figure 2: Identifying "Scopes" for Web Resources

Figure 2 details the process of assigning scopes to Web resources. Our domain knowledge is organized in a common Geographic Knowledge Base (GKB), which integrates the information that we have on named entities and their location attributes. Geographic augmenters then transform PWDs into SAWDs using a sequence of steps involving lexical analysis, geographic entities recognition and, finally, geographic scope inference. The following processing phases take place in generating SAWDs:

GKB Construction: The GKB maintains the identified associations between geographic entities contained in Web pages, and the scopes and entities described in imported ontologies. Examples of the former include addresses and landmarks, postal codes and fixed phone numbers. Scopes are defined by the system designers, and correspond to administrative regions or popular districts. Geographically located entities are obtained from external information sources, such as WHOIS and DNS registrars, the Portuguese postal codes database, and directories of organizations.

Lexical Analysis: The text contained in Web documents is divided into lexical tokens (words or field markers). We also consider possible parts-of-speech.

Geographic Entity Recognition: Entities are identified by pattern matching against the GKB, and using parts-of-speech, syntactic and orthographic features (ex: capitalization). A recent study showed that geographical name extraction with a small gazetteer can produce good results [20]. The disambiguation process is needed to distinguish between "Braga" as city and "Braga" as a person surname, for example. This may require pattern analysis (expressions like "city of" or "located in" are common before geographical entities), and concept recognition within compound terms (that is, recognizing "City of Braga" in spite of "Mr. João Braga"). In addition, we expect ontologies to help with the disambiguation process, for instance by sharing relevant hyperonyms.

Web Inference: One of the ideas behind most link structure analysis [7] is the topic locality assumption [8]: content and hypertext links are correlated, i.e., a link from document A to document B means that documents A and B are on the same topic. We are planning on propagating information from pages with a known geographical scope, to those that are linked to/from it, in order to infer scopes for Web pages, or to increase the confidence level on the assignment process. This is similar to the work presented in [17]. If there is a connection between pages A and B, and if the geographic scope of page A was already identified, than we can say page B is, to some extent, relevant to the same scope.

Semantic Inference: Contents published on the Web reflect the dinamycs of our communities. Sites are continuously changing and we need to identify the scopes of previously unregistered entities. The scopes of pages and the associations between scopes and entities are also permanently changing. The names of local football players, when referenced, are relevant to the supporters of their area. However, when they are transfered to another team, the association progressively vanishes. The association between unknown entities and scopes can be obtained using probabilistic methods and existing knowledge: if names M,N are frequent in a set of Web documents with scope S and M corresponds to scope S, then N can also be associated to scope S. In MindNet [23], an inference procedure using similarity measures allows the identification of previously unknown semantic relations. We can also use previously proposed similarity measures [24] to generate topic maps with the discovered knowledge.

One difficulty with geographic entity recognition is on the decision of whether an entity has geographic significance, when it is composed of geographic terms. If "Lisbon" is generally considered a geographic entity, the expression "Mayor of Lisbon" is not unanimously associated with geographic semantics. Our decisions for considering names as geographic entities will be based on rules generated by analysis of the consensus among the participants on a named entity recognition evaluation contest [22]. We are promoting this contest in association with Linguateca, a distributed language resource center for Portuguese. In this evaluation contest, documents will be manually parsed by a group of annotators. The rules and heuristics will derive from the agreements and objections recorded while marking-up the test collection provide the basis for determining the geographic relevance of extracted names when performing pattern matching against our knowledge base.

3. GEOGRAPHIC RETRIEVAL

Having "geographical scopes" assigned to Web resources, we can exploit their use for IR in various ways. The tumba! search engine already supports inter-document similarity retrieval techniques based on related pages and result sets clustering [18]. With geographic scopes information, similarity can also be computed in

terms of geographic relatedness: two documents are similar if they relate to the same or nearby geographical scopes. Retrieval methods that can then make use of this notions include: i) retrieving documents that describe resources nearby; and ii) clustering documents according to their geographical scopes.

We have already observed that users frequently input geographic entities in their queries. In addition, they may also indirectly provide their geographic position when accessing our search service [12]. With this information in hand, we can both derive names of geographic entities to further restrict matching result sets and compute similarities based on the distance between the perceived user location and the geographic scope of Web resources [4].

4. RELATED WORK

The WebFountain project is an example of a computer cluster designed to analyze massive amounts of textual information, enabling the discovery of trends, patterns and relationships [11]. The architecture integrates applications that focus on specific tasks, using multi-disciplinary text analytic approaches to extract data from Web resources.

The SPIRIT project aims to develop a search engine aware of geographical terminology, using ontologies [13]. Our framework differs on the emphasis put on geographic named entities recognition, the use of linguistic methods, and the automatic generation of ontologies associating entities to geographic scopes.

Previously, the NetGeo project also concerned geographic locations in the context of the Internet, collating information from multiple sources in order to assign the most probable longitude/latitude to IP addresses [21].

5. CONCLUSIONS

We presented our approach for automatically identifying geographic scopes in Web pages. A shared knowledge base is used to augment RDF-based descriptions of crawled Web pages with geographic meta-data. This work is part of a larger project which will also involve the creation innovative IR algorithms in our Web search engine, using the notion of "geographical relatedness".

6. REFERENCES

[1] W3c semantic web. `http://www.w3.org/2001/sw/`.

[2] Xmlbase project homepage. `http://xldb.fc.ul.pt/index.php?page=XMLBase`.

[3] Iso/iec 13250:2000 topic maps: Information technology – document description and markup languages, 1999.

[4] A. P. Afonso. *Contribuições Metodológicas para o Desenvolvimento de Assistentes de Informação Personalizada*. PhD thesis, Department of Informatics, University of Lisbon, May 2004. DI/FCUL TR-04-3.

[5] A. Arasu, J. Cho, H. Garcia-Molina, A. Paepcke, and S. Raghavan. Searching the Web. *ACM Transactions on Internet Technology*, 1(1):2–43, August 2001.

[6] N. Cardoso and M. J. Silva. A statistical study on the tumba! search engine query logs, 2004. (To Appear).

[7] S. Chakrabarti, B. E. Dom, S. R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. Kleinberg. Mining the Web's link structure. *Computer*, 32(8):60–67, August 1999.

[8] B. D. Davison. Topical locality in the Web. In *Proceedings of SIGIR-00, the 23rd ACM International Conference on Research and Development in Information Retrieval*, pages 272–279, 2000.

[9] D. Gomes and M. J. Silva. A characterization of the Portuguese Web. In *Proceedings of the 3rd ECDL Workshop on Web Archives*, Trondheim, Norway, August 2003.

[10] R. Grishman. Information extraction: Techniques and challenges. In M. T. Pazienza, editor, *Lecture Notes In Artificial Intelligence*, volume 1299, pages 10–27. Springer-Verlag, 1997.

[11] D. Gruhl, L. Chavet, D. Gibson, J. Meyer, P. Pattanayak, A. Tomkins, and J. Zien. How to build a webfountain: An architecture for very large-scale text analytics. *IBM Systems Journal - Utility Computing*, 43(1), 2004.

[12] J. Hjelm. *Creating Location Services for the Wireless Web*. John Wiley & Sons, 2002.

[13] C. B. Jones, R. Purves, A. Ruas, M. Sanderson, M. Sester, M. van Kreveld, and R. Weibel. Spatial information retrieval and geographical ontologies: An overview of the spirit project. In *Proceedings of SIGIR-02, the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 387–388. ACM Press, August 2002.

[14] H. Katz, D. Chamberlin, D. Draper, M. Fernandez, M. Kay, J. Robie, M. Rys, J. Simeon, J. Tivy, and P. Wadler. *XQuery from the experts : A Guide to the W3C XML Query Language*. Addison-Wesley, 2003.

[15] R. Kosala and H. Blockeel. Web mining research: A survey. *SIGKDD Explorations: Newsletter of the ACM Special Interest Group (SIG) on Knowledge Discovery and Data Mining*, 2, 2000.

[16] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, 1999.

[17] M. Marchiori. The limits of web metadata, and beyond. In *Proceedings of WWW-98, the 7th International World Wide Web Conference*, April 1998.

[18] B. Martins. Inter-document similarity in web searches, 2004. (To Appear).

[19] B. Martins and M. J. Silva. A statistical study of the wpt-03 corpus. DI/FCUL TR 04–04, Department of Informatics, University of Lisbon, May 2004.

[20] A. Mikheev, M. Moens, and C. Grover. Named entity recognition without gazetteers. 1999.

[21] D. Moore, R. Periakaruppan, and J. Donohoe. Where in the world is netgeo.caida.org? In *Proceedings of INET-2000, The 10th Annual Internet Society Conference*, July 2000.

[22] C. Mota, D. Santos, and E. Ranchhod. Avaliação de Reconhecimento de Entidades Mencionadas: princípio de AREM. in Diana Santos (ed.), Avaliação Conjunta: um novo paradigma no processamento computacional da língua portuguesa, forthcoming.

[23] S. D. Richardson. *Determining Similarity and the Inferring Relations in a Lexical Knowledge-Base*. PhD thesis, 1997.

[24] P. P. Senellart and V. D. Blondel. Automatic discovery of similar words. In M. W. Berry, editor, *A Comprehensive Survey of Text Mining*. Springer-Verlag, 2003.

[25] M. J. Silva. The case for a portuguese web search engine. In *Proceedings of ICWI-2003, the IADIS International Conference WWW/Internet 2003*, November 2003.