Spatial Search Engines

Cláudio Elízio Calazans Campelo
University of Campina Grande, Brazil

Cláudio de Souza Baptista
University of Campina Grande, Brazil

Ricardo Madeira Fernandes
University of Campina Grande, Brazil

INTRODUCTION

It is well known that documents available on the Web are extremely heterogeneous in several aspects, such as the use of various idioms, different formats to represent the contents, besides other external factors like source reputation, refresh frequency, and so forth (Page & Brin, 1998). Altogether, these factors increase the complexity of Web information retrieval systems.

Superficially, traditional search engines available on the Web nowadays consist of retrieving documents that contain keywords informed by users. Nevertheless, among the variety of search possibilities, it is evident that the user needs a process that involves more sophisticated analysis; for example, temporal or spatial contextualization might be considered. In these keyword-based search engines, for instance, a Web page containing the phrase "…due to the company arrival in London, a thousand java programming jobs will be open…" would not be found if the submitted search was "jobs programming England," unless the word "England" appeared in another phrase of the page. The explanation to this fact is that the term "London" is treated merely like another word, instead of regarding its geographical position. In a spatial search engine, the expected behavior would be to return the page described in the previous example, since the system shall have information indicating that the term "London" refers to a city located in a country referred to by the term "England." This result could only be feasible in a traditional search engine if the user repeatedly submitted searches for all possible England sub-regions (e.g., cities). In accordance with the example, it is reasonable that for several user searches, the most interesting results are those related to certain geographical regions. A variety of features extraction and automatic document classification techniques have been proposed, however, acquiring Web-page geographical features involves some peculiar complexities, such as ambiguity (e.g., many places with the same name, various names for a single place, things with place names, etc.). Moreover, a Web page can refer to a place that contains or is contained by the one informed in the user query, which implies knowing the different region topologies used by the system.

Many features related to geographical context can be added to the process of elaborating relevance ranking for returned documents. For example, a document can be more relevant than another one if its content refers to a place closer to the user location. Nonetheless, in spatial search engines, there are more complex issues to be considered because of the spatial dimension concerning on ranking elaboration. Jones, Alani, and Tudhope (2001) propose a combination of Euclidian distance between place centroids with hierarchical distances in order to generate a hybrid spatial distance that may be used in the relevance ranking elaboration of returned documents. Further important issues are the indexing mechanisms and query processing. In general, these solutions try to combine well-known textual indexing techniques (e.g., inverted files) with spatial indexing mechanisms. On the subject of user interface, spatial search engines are more complex, because users need to choose regions of interest, as well as possible spatial relationships, in addition to keywords. To visualize the results, it is pleasant to use digital map resources besides textual information.

BACKGROUND

Numerous contributions have been made in the information retrieval (IR) area since 1960's decade. Nevertheless, due to Web continuous growth, research in this field is still in infancy.

Baeza and Ribeiro (1999) say that IR brings some challenges, such as how to determine the real user needs, as well as supply their expectations through relevant document subsets. In IR systems, it is necessary to analyze both semantics and syntax of document contents, which may return imprecise results. According to Kowalsky (1997), the aim of an IR system is to minimize the overhead of finding the expected information. Classical IR models consider that a document is described by a set of indexed terms. Some of these models

also take into account different terms importance at the same document. This importance is called weight (w), and can be represented by a numeric value. The most well-known classical models are Boolean, probabilistic and vector. The vector one, proposed by Gerard Salton, has a greater acceptance between researchers and is the most utilized in current IR applications.

The Web brings new features and difficulties to the IR process, due to both the heterogeneity of the underlying documents and the approaches used to present them. Studies have demonstrated that a large amount of information disposed on the Internet has some kind of geographical context. For instance, the locale where the information was created, the referenced information locale, the place where most information consumers live, and so forth. However, traditional search engines do not consider this spatial context in their information organization and retrieval process.

The requirement for efficient information supported by the knowledge about a specific domain raised the concern on developing ontologies that model many associated concepts. Concerning geographical IR, the use of ontologies can be extremely important for geographical features representation of documents. Fu, Jones, and Abdelmoty (2005) suggest the existence of a primary ontological component, place ontology, that provides the terminology and geographical space structure modeling. Such ontology has a fundamental role, for instance, in user query interpretation, relevance ranking elaboration, and metadata extraction.

Surely, the relevance ranking elaboration is one of the main processes in a search engine, since it is directly related to user interest. In traditional systems, the ranking can be produced through a variety of techniques, for example, similarity measures between query and returned documents using the spatial-vector model (Baeza & Ribeiro, 1999). Currently, one of the most accepted methods is the PageRank (Page & Brin, 1999) that uses the Internet link structure to produce its ranking.

Research on spatial search engines is very incipient. It has addressed some information retrieval subareas aiming to develop efficient data structures and algorithms for space-textual indexing; to elaborate efficient approaches for relevance ranking using the geographic context; and to detect and model the geographic scope.

There are different approaches to extract geographic information from Web-crawled documents. Buyukkokten, Cho, Molina, Gravano, and Shivakumar (1999) associate domain name IP addresses to telephone code area, and by using postal code of the Web site, they enable to match place names to geographic coordinates. McCurley (2001) introduced the geocoding concept, which enables one to associate geographic coordinates to Web pages. McCurley also has used several terms that are useful to geocode a Web page, for instance, postal code, city names, and telephone numbers. Nonetheless, there is no discussion on techniques to extract and eliminate ambiguity. Gravano (Gravano, Hatzivassiloglou, & Lichtenstein, 2003) has implemented automatic geocoding.

More recently, some research projects have presented relevant results in this field, as, for example, the GeoTumba one (Chaves, Silva, & Martins, 2005). GeoTumba contains a repository based on a domain-independent metamodel to integrate geographical knowledge collected from several sources. Silva, Martins, Chaves, Afonso, and Cardoso (2006) focus on techniques for geographical features extraction from large collections of Web documents by using a method that involves the attribution of geographic scope through the GraphRank algorithm, which is inspired in the PageRank one (Page & Brin 1999). Silva et al. (2006) show three sets of heuristics used in the process of georeferencing Web pages.

Another important research project is the SPIRIT (*spatially-aware information retrieval on the Internet*), which focuses on geographic information retrieval and issues involving the semantic Web. This project proposes a multimodal interface that provides text and maps; spatially aware ontologies; query expansion and relevance ranking based on geographic ontologies; spatial indexes for the document collection; and a learning mechanism for extracting geographic context from Web documents that generates spatial metadata. An overview of this project can be found in Jones, Abdelmoty, Finch, Fu, and Vaid, 2004), which addresses the architecture, indexing mechanisms, and spatial ontologies.

Markowetz, Chen, Suel, Long, and Seeger (2005) use the geocoding concept divided into three steps: *geoextraction*, *geomatching and ge propagation*. Their work is inspired in Buyukkokten et al. (1999), and Ding, Gravano, and Shivakumar (2000). Markowetz, Brinkhoff, and Seeger (2004) propose a relevance ranking that may balance between text and spatial ranking.

TOWARD A SPATIAL SEARCH ENGINE

This section presents the GeoSEn, geographic search engine, project that has been developed in our laboratory, which may be accessed at http://www.lsi.dsc.ufcg.edu.br/geosen. GeoSEn is a Web spatial information retrieval system that uses geographical scope detection mechanisms (set of places that Web element contents can be associate) to better index Web documents.

Some studies have proposed a Web geographical information retrieval system. The major concern of these prototypes is to detect pages spatial features and represent them, providing documents retrieval according to a relevance ranking, elaborated considering the geographical location of the documents. However, the presented mechanisms for geographical scope detection and relevance ranking elaboration still have limitations and demand for innovative contributions.