Geographically-Aware Information Retrieval on the Web

I

Claudio E.C. Campelo
Federal University of Campina Grande, Brazil

INTRODUCTION

Numerous contributions have been made in the field of *Information Retrieval (IR)* since the 60s to maximise the amount of relevant information which can be retrieved from large collections of information resources. However, with the rapid growth of the World Wide Web (WWW) since the 90s, additional challenges have been posed to the IR task. Moreover, the Web users have been changing gradually the way they use the Web. People are now connected to each other via social networks and are using mobile devices intensively. It has led to considerable modification in the way the users look for information and evaluate their relevance. This new scenario has motivated the development of *specialised Web search engines* which aim to meet particular user needs.

Extremely roughly, traditional Web search engine algorithms are based on keyword matching, which imposes a number of limitations to perform more specialised IR tasks, such as the retrieval of geographic information. For example, a Web document containing the sentence "...with the establishment of the company in *Leeds*, thousands of medical equipments will be manufactured daily..." would not be retrieved by a typical search engine by querying the system using the keywords "medical equipment manufacturer england," since no word in the document's text would match the term 'england' (unless, of course, some other text fragment in the document also contains the word 'England'). This happens because the expression 'england' is treated as an ordinary term, not as a geographic place. In a *geographically-aware search engine* this document should be returned, since the system should be able to infer that the term 'Leeds' refers to a city which is located in a country which can be referred to by the term 'england'.

Geographic Information Retrieval (GIR) is a recent research area which has become notably attractive. Geographic Web search engines are specialisations of standard Web search engines, adding to them the ability to identify geographic contexts of Web resources (e.g., texts, images, movies) and to index them according to such contexts. This article presents the main topics of research within GIR and overviews significant contributions in the field.

BACKGROUND

Most of the information available on the Web has some sort of *geographic context*. This context includes the place where the information has been created, places referred to in the document, and those in which the information is regarded as more relevant. Some experiments reported in the literature have demonstrated that a considerable amount of Web pages contains terms which may be derived into geographic places. Examples of such terms include place names, telephone numbers and postcodes.

McCurley (2001) observed that approximately 8.5% of Web pages contain a telephone number, 4.5% contain a postcode, and 9.5% contain one of the two. Silva et al. (2006) analysed 3,775,611 Web pages and noticed that they contain an average of 2.2% references to Portuguese cities (obviously, if other place names apart from cities had been taken into account, such numbers are likely to be even greater). Nevertheless, traditional information retrieval systems do not consider this context in their parsing, indexing and retrieval process. This section presents the main motivations for the development of GIR. For an appropriate understanding of the field, it is categorised here into different sub-areas and the major challenges and problems faced in each of them are described.

DOI: 10.4018/978-1-4666-5888-2.ch383

Geographic References Detection

Web documents are collected from the Web by a robot known as Web crawler or Web bot. Then these documents are usually submitted to a full-text parsing process, where the terms which constitute the documents are identified and used to index these documents for fast retrieval. In a geographic search engine, the parsing process includes an additional functionality, which consists of identifying any term which can be mapped to an existing geographic place. These terms can be, for example, city names, postcodes and phone numbers. Here, these terms are referred to as *geographic references*.

One of the main challenges encountered in this sub-area of GIR is dealing with the *ambiguity* which affects geographic terms. For instance, there are different places with identical names (e.g., London in the UK and London in Canada); places with people's names (e.g., Charlotte, a usual female first name which is also the name of a city in the U.S. state of North Carolina); and identical terms which may refer to either places or things (e.g., Bath, which may refer to the container which holds water for bathing or to the city in the county of Somerset in South West England). Therefore, a GIR system should be able to disambiguate *candidate terms* to determine whether they are actually referring to a geographic place (i.e. whether the term is a *valid geographic reference*) and, if so, to determine which particular instance of that geographic name is intended.

Geographic Scope Definition

The *geographic scope* of a Web resource can be defined as the set of places they are associated with. The term is often used interchangeably with *geographic context*. Once the valid geographic references are identified for a given Web document, they are used as an input of specialised algorithms which models the document's geographic scope. The main challenge in this sub-area of GIR is to define the most appropriate representation of the documents' geographic scope in order to support the process of indexing and search. For example, the geographic scope can be described in terms of the content of the document or in terms of the place where the information has been created. To illustrate, a German electronic newspaper may publish news about a fact occurred in London, UK. Buscaldi (2011) refers to the place where the information has been created as the 'source' of information, and argues that it is an important feature, specially for local collections (e.g., local newspapers). This distinction may be useful to amplify the ways geographic information can be retrieved. The literature also distinguishes the geographic scope as *simple* or *multiple*, in terms of whether a Web resource is associated with a single or with multiple localities, respectively. The geographic scope representation may also vary in relation to the way they are stored at the data level, such as using place names, or a single geographic coordinate per locality (e.g., the centroid) or a set of geographic coordinates (e.g., a polygon), amongst others. The geometrical representation of a document's geographic scope is known as the document's *spatial footprint*.

Spatio-Temporal Indexing

Indexes are widely used in IR systems to promote high velocity in the search process when handling a large collection of documents (Baeza-Yates & Ribeiro-Neto, 1999). The mechanisms for indexing in Web search engines are based on techniques similar to the traditional IR systems, usually based on *inverted files*. Each document is converted into a set of term occurrences called hits. These hits store information about the position of the word within the document, besides some presentation information such as font size and use of bold. In addition, the system stores information on anchors and links.

Additional difficulties are encountered for indexing in geographic search engines, since both textual and spatial information must be taken into account to retrieve Web resources efficiently. These hybrid indexes are often referred to as *spatio-textual indexes*. Once the geographic references present in a document are identified and its geographic scope is defined, the document must be indexed by the set of places they are associated with. Using textual indexes for both textual and spatial data would mean employing the same structure used in inverted files of traditional search systems, where it is necessary that the argument specified in the user query matches the term present in the indexed document. However, for handling spatial information, this assumption is not admissible. For example, the user may specify a region which contains a certain place found in a document, but may not specify such a place