Engineering Applications of Artificial Intelligence 51 (2016) 202–211

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Geographical localization of web domains and organization addresses recognition by employing natural language processing, Pattern Matching and clustering

Paolo Nesi [1], Gianni Pantaleo [1], Marco Tenti [1]

DISIT Lab, Department of Information Engineering (DINFO), University of Florence, Via S. Marta 3, 50139 Firenze, Italy

ARTICLE INFO

Available online 29 January 2016

Keywords:
Geographic Information Retrieval
Geoparsing
Geocoding
Data mining
Natural language processing
Hierarchical Clustering

ABSTRACT

Nowadays, the World Wide Web is growing at increasing rate and speed, and consequently the online available resources populating Internet represent a large source of knowledge for various business and research interests. For instance, over the past years, increasing attention has been focused on retrieving information related to geographical location of places and entities, which is largely contained in web pages and documents. However, such resources are represented in a wide variety of generally unstructured formats, and this actually does not help final users to find desired information items. The automatic annotation and comprehension of toponyms, location names and addresses (at different resolution and granularity levels) can deliver significant benefits for the whole web community by improving search engines filtering capabilities and intelligent data mining systems. The present paper addresses the problem of gathering geographical information from unstructured text in web pages and documents. In the specific, the proposed method aims at extracting geographical location (at street number resolution) of commercial companies and services, by annotating geo-related information from their web domains. The annotation process is based on Natural Language Processing (NLP) techniques for text comprehension, and relies on Pattern Matching and Hierarchical Cluster Analysis for recognizing and disambiguating geographical entities. Geotagging performances have been assessed by evaluating Precision, Recall and F-Measure of the proposed system output (represented in form of semantic RDF triples) against both a geo-annotated reference database and a semantic Smart City repository.

1. Introduction

In March 2015 the number of online active Web sites had been estimated in about 900 million (878 million, as reported by Netcraft,[2] 903 million according to the CIA World Factbook,[3] 930 million as stated by Internet Live Stats[4]). As a global leader in domain names and internet security, Verisign[5] periodically reviews the state of web domain name industry, reporting that the fourth quarter of 2014 ended with a total of 288 million Top-Level domain name registrations (TPDs), with an increase of 1.3 percent (about 4 million domain names) over the third quarter of 2014.

Such a huge amount of web resources represents an extremely vast source of knowledge, for the most part embedded in the textual content of web pages and documents (in the following, the term Web page will be used to describe every web resource identified by its own unique URL, containing a textual content that can be navigated and parsed, while the term Web document will be used to identify a larger variety of web available text files - such as .doc, .rtf, and .pdf – whose textual content can be downloaded and parsed). However, it is becoming increasingly difficult, for final users, to extract specific information items of interest, since web resources are not yet fully structured in a machine-readable format. The majority of web pages and documents are still a collection of unstructured text formats without any explicit meaning automatically inferable by machines (Schmidt et al., 2013a). An

E-mail addresses: paolo.nesi@unifi.it (P. Nesi), gianni.pantaleo@unifi.it (G. Pantaleo).

[1] http://www.disit.dinfo.unifi.it
[2] http://news.netcraft.com/archives/category/web-server-survey/
[3] https://www.cia.gov/library/publications/the-world-factbook/
[4] http://www.internetlivestats.com/total-number-of-websites/
[5] http://www.verisigninc.com/en_US/innovation/dnib/index.xhtml

http://dx.doi.org/10.1016/j.engappai.2016.01.011
0952-1976/

early attempt for adding structure to HTML pages were Microformats.[6] Microformats, developed as part of the HTML5 standardization efforts, defines fixed vocabularies to annotate specific entities such as people, relationships, organizations and places, calendar entries, products, and cooking recipes. within HTML pages (Loglisci et al., 2012). A more recent approach is represented by the Semantic Web applications: an increasing number of HTML pages embed structured XML/RDF data and schemas, according to the RDFa format (Bizer et al., 2008); besides, dedicated Ontologies and Taxonomies have been developed, such as the GoodRelations Ontology, which semantically models and describes details of business services and products in a fully integrated way with the schema.org markup vocabulary, used by the most important search engines such as Google, Yahoo!, Bing/Microsoft, and Yandex mobile applications (Hepp, 2013). Recently, a growing interest has arisen in the field of semantic data enrichment, since it covers the representational needs and interoperability requirements within the expanding e-commerce scenarios. However, since all these represent emerging standards, the vast majority of actual available online resources do not support yet such new reference benchmarks. Therefore, strong interests and needs are perceived to extract structured information in web pages, in a large variety of fields and application areas. Until the semantic enrichment of web content will not reach a significant degree of penetration, automatic annotation of information items represents an alternative to manual annotation, which is an extremely inefficient and time-consuming process.

This paper focuses on the annotation of geographic information from unstructured text in web pages and documents, with the aim of extracting geographical location of commercial entities. This process is defined as Geographic Information Retrieval (GIR), as it concerns the extraction of information involving some kind of geospatial reference (Mandl et al., 2008). Understanding place names mentioned in textual data can provide great benefits for data mining systems and search engines, enhancing the capabilities of geographic-based queries and filtering. From past studies emerged that 15% of the queries submitted to search engines contain references to geographic names (Anderson and Kohler, 2004). Analyzing some quite recent reports, in May 2011, 23% of USA citizens used location-based services. This number increased to 41% in February 2012, and it can be assumed that such a trend is still growing (Zickuhur, 2012). Automatically retrieving address information about business entities, associations and organizations is an aspect that attracts a lot of commercial and research interest, ranging from geo-marketing to criminal investigation and fraud detection. Location-based applications can take advantage from automated harvesting of address data from Web sites, for instances recommendation-based systems can provide spatial-based suggestions on the surrounding of a user. An important application area for geographical annotation technology is found in the recently developing Smart City frameworks, aiming at helping citizens by providing different services and useful information on public available Open Data (OD), including geographical information and spatial location of places of interest, real-time traffic and parking structures, as well as any other kind of municipality resource which can be geolocated. This kind of services requires a very high spatial resolution (generally at street number level), although this is still far to be achieved by present GIR systems with a sufficient degree of confidence and reliability. The solution to this issue is not trivial, and it concerns not only with the improvement of the current GIR tools and services, but also with a desirable standardization of OD formats, descriptors and languages. Smart City services are often based upon unstructured Open Data representing public administration services and utilities which are not always geographically referenced. For this purpose, and moreover to cover commercial stakeholders which are not included in public OD, additional information is needed. Such a requirement is met by the presence of many online web domains that are usually representative of activities undertaken by business organizations. Consequently, the necessity arises to associate commercial web domains to geographic information related to their corresponding physical entities.

The main objective of this work is to present a system for extracting administrative information, including addresses and geographical coordinates, of web-visible human activities (intended, in the most generic meaning, all those human activities – ranging from commercial and research organizations, private companies and services, and Public Administration services – which are associated to a public available web domain) from unstructured text contents hosted in their web sites. The main areas and aspects of Engineering Applications of Artificial Intelligence addressed in the paper are Data Mining and NLP, in the specific: Text Mining and Annotation, Named Entity Recognition, Part-of-Speech (POS) tagging. Evaluation experiments have been conducted focusing on mining web domains and pages owned by organizations located in the Tuscany region, Italy. Actually, according to the Italian National Institute of Statistics (ISTAT), Tuscany is one of the Italian regions with the highest number of firms and companies (reported as a total of 330 thousands registered companies in the last 2011 census, which is equivalent to nearly 80 firms per thousand citizens). However, only a small percentage of such commercial operators is estimated to be actively registered to the Open Data repositories provided by regional Public Administrations, as well as to external resources (e.g.: the GoodRelations users community). The proposed framework aims at filling this gap. The present paper is organized as follows: Section 2 illustrates the related work, in terms of state of the art and research issues for both commercial and research literature; in Section 3, the functional architecture of the proposed system is presented; in Section 4, a two-steps validation of the system is reported (in order to assess both address information and geographic coordinates extraction); finally, Section 5 is left for conclusions and future perspectives.

2. Related work

The processes of recognizing geographic context and assigning spatial coordinates are commonly referred to as *geoparsing* and *geocoding*, respectively (Scharl, 2007). Geoparsing deals with parsing unstructured text and extracting keywords and keyphrases describing geographical references, including the extraction of terms and/or metadata representing physical, natural and human-made features (e.g.: countries, cities, roads, addresses, postal codes, telephone numbers, and buildings, but also forests, rivers, lakes, and mountains). In its most general meaning, geoparsing refers to the extraction of toponyms in texts (Pouliquen et al., 2004), and it is related to Natural Language Processing (NLP) and connected tasks such as Named Entity Recognition (NER, addressing the detection of general named entities) and Word Sense Disambiguation (WSD, aiming at resolving ambiguities occurring in presence of polysemous words and expressions). Geocoding, on the other hand, is defined as the process of mapping geographical annotations to their real-world counterparts by associating spatial coordinates (latitude, longitude and, in case, altitude).

[6] http://microformats.org/