Spatially Aware Term Selection for Geotagging

Geetha.G
Assistant Professor, Department of Information Technology,
VR Siddhartha Engineering College, Vijayawada,
Andhra Pradesh, India,

Madhavi katamaneni
Assistant Professor, Department of Information Technology,
VR Siddhartha Engineering College, Vijayawada,
Andhra Pradesh, India,

S.Sunitha
Assistant Professor, Department of Information Technology,
VR Siddhartha Engineering College, Vijayawada,
Andhra Pradesh, India,

Abstract- This paper presents methods used to extract geospatial information from web pages for use in SPIRIT, a new Geographic Information Retrieval (GIR) system for the web. The resulting geospatial markup tools have been used to annotate around 900,000 web pages taken from a 1TB web crawl, focused on regions in the UK, France, Germany and Switzerland. This paper discusses a versatile geo-parsing tool for extracting spatial metadata based upon the GATE Information Extraction (IE) system, and a simple geo-coding program based on default sense to assign spatial coordinates to extracted locations. A preliminary analysis of markup accuracy for geo-parsing and geo-coding is provided, and an initial statistical and geographical analysis of the SPIRIT collection presented

Keywords:Geographic information retrieval(GIR),Information extraction(IR).

I. INTRODUCTION

Many documents on the web contain geospatial information including addresses, postal codes, hyperlinks and geographic references [1][2]. This information can be exploited and used to provide spatial awareness to information systems. These include transport timetables, routing systems for motorists, map-based web sites and location-based services (e.g. Google Local and Yellow Pages). A key part of providing such services is theextraction and use of geospatial information. In this paper we discuss approaches used in the Spatially-Aware Information Retrieval on the Internet (SPIRIT) project (http://www.geo-spirit.org/) to generate a sample web document collection for prototyping a working GIR system [3]. Extracting geospatial references from documents involves two main tasks: identifying geographic references and assigning them spatial coordinates. These are commonly referred to as geo-parsing and geo-coding respectively [4].

1.1 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential

1.2 ECONOMIC FEASIBILIY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

1.3 TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

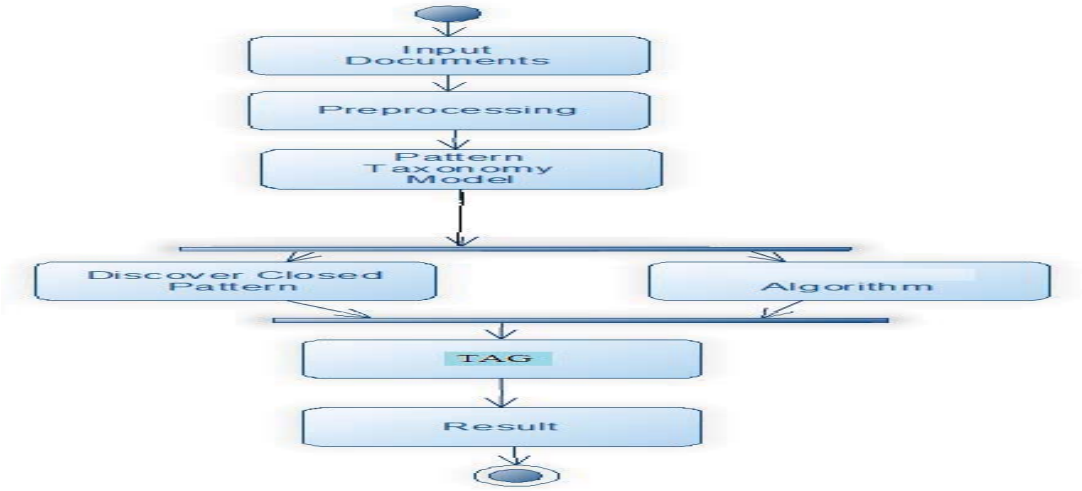
II.LITERATURE SURVEY

Many types of text representations have been proposed in the past. A well known one is the bag of words that uses keywords (terms) as elements in the vector of the feature space. In [21], the tf*idf weighting scheme is used for text representation in Rocchio classifiers. In addition to TFIDF, the global IDF and entropy weighting scheme is proposed in [9] and improves performance by an average of 30 percent. Various weighting schemes for the bag of words representation approach were given in [1], [14], [38]. The problem of the bag of words approach is how to select a limited number of features among an enormous set of words or terms in order to increase the system's efficiency and avoid overfitting [41]. In order to reduce the number of features, many dimensionality reduction approaches have been conducted by the use of feature selection techniques, such as Information Gain, Mutual Information, Chi-Square, Odds ratio, and so on.

III.PROPOSED METHOD

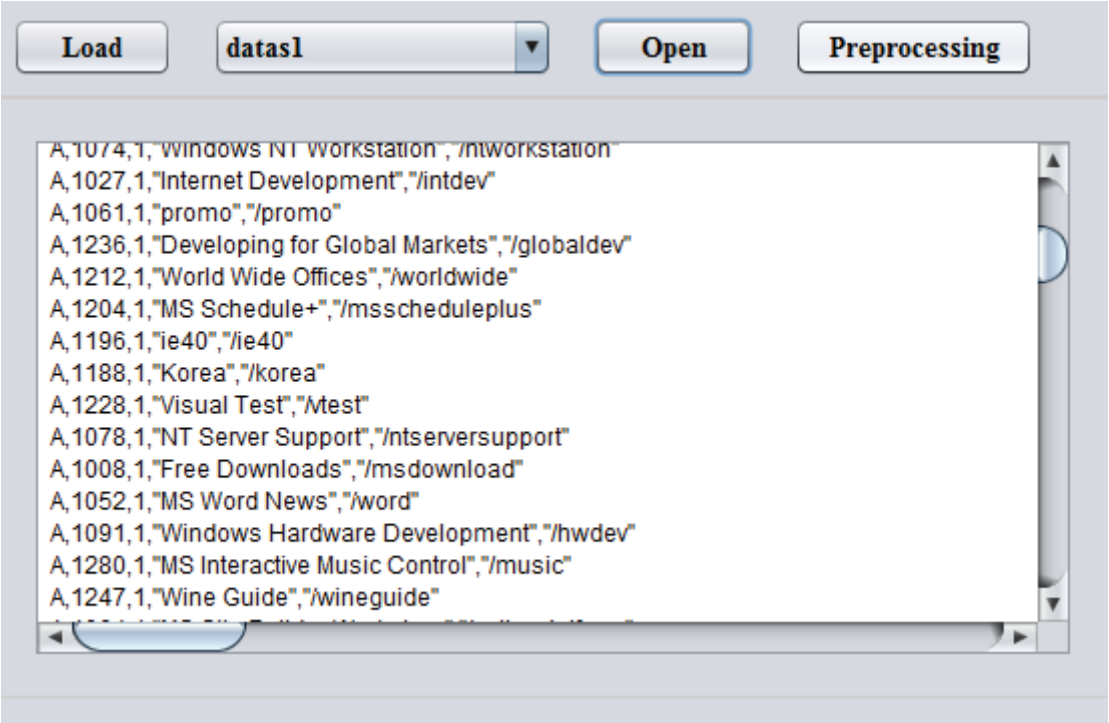
SPIRIT (SPATIALLY-AWARE INFORMATION RETRIEVAL ON THE INTERNET)

This paper we discuss approaches used in the Spatially-Aware Information Retrieval on the Internet (SPIRIT) project (http://www.geo-spirit.org/) to generate a sample web document collection for prototyping a working GIR system. Extracting geospatial references from documents involves two main tasks: identifying geographic references and assigning them spatial coordinates. These are commonly referred to as geo-parsing and geo-coding respectively



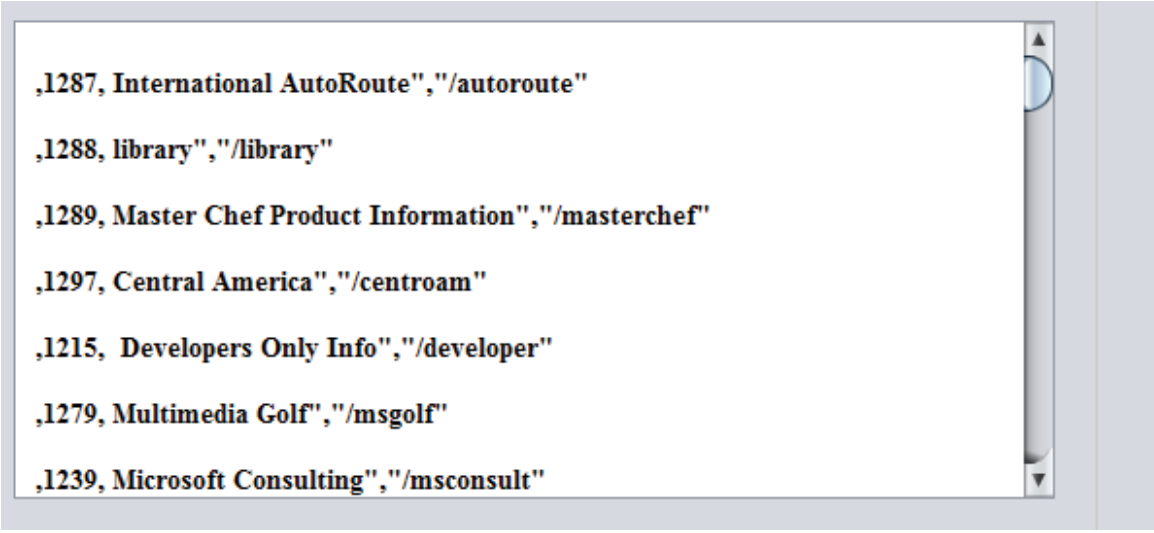
IV.RESULTS AND OBSERVATIONS

4.1 [COLLECTED DATA]



4.2 [PREPROCESSING]

REMOVING STOP
WORDS

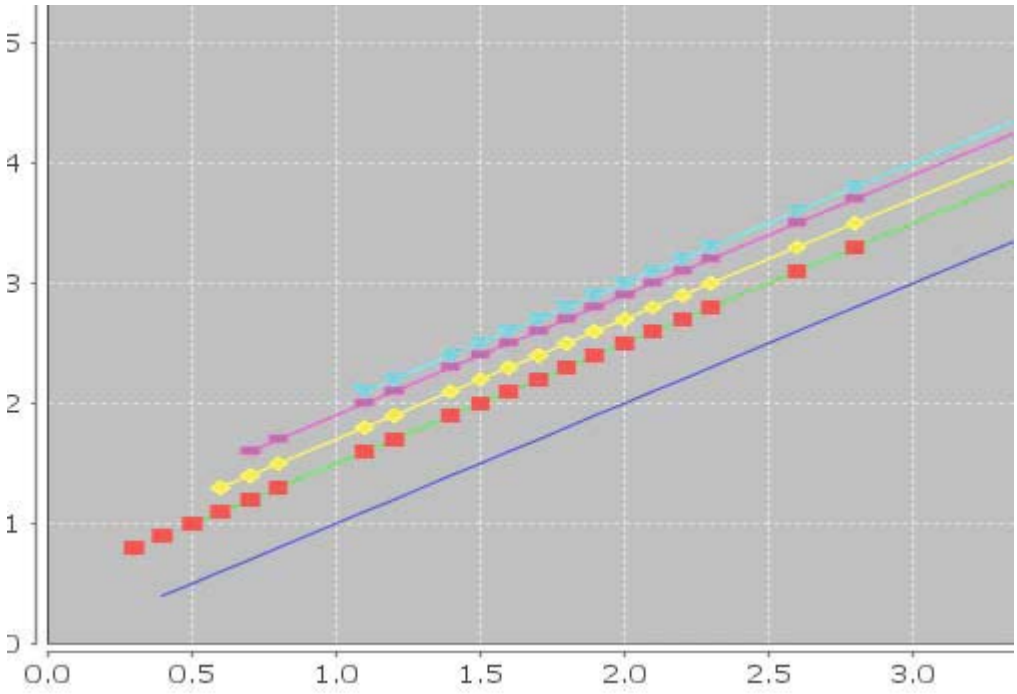


4.3 [STEMMING]

4.4 [T-SUPPORT]

worldwide	2.1	
officefreestuff"		2.3
kb"	0.6	
sql"	1.2	
msgarden"	0.2	
merchant"	0.4	
encarta"	0.3	
advtech"	0.6	
exchange"	0.5	
chile"	0.3	
msscheduleplus"		0.5
k-12"	1.1	
china"	0.3	
imedia"	1.6	
logostore"	0.6	
vbscripts"	0.7	
feedback"	0.6	
colombia"	0.3	
canada"	0.3	
southafrica"	0.5	
services"	2.0	

4.5 [RESULT]



V.CONCLUSIONS

This paper presents work from the SPIRIT project on extracting spatial metadata from web pages for prototyping a working GIR system . Simple methods for geo-parsing and geo-coding have been presented which address specific constraints including execution time and language independence. Methods used are relatively simple and provide a baseline upon which to construct more complex approaches. In particular, the geo-parsing method used has been based on the GATE system providing a versatile framework in which to develop custom tools. The use of gazetteer lookup provides an F1 score of 0.7148 when used with removal of commonly occurring words (stopwords) and other entities including person names to address cases when gazetteer entries are used in a non-geographical sense. The geo-coding method is also simple based on default sense and matches between metadata provided by the geographical resources and local context. Using a small set of manually annotated data, 89% of locations are correctly assigned a unique identifier (UID). In particular, the grounding method also deals with resource selection in addition to referent ambiguity. Whether this accuracy of markup is sufficient in practice is still being investigated, but both user and system evaluation of the SPIRIT prototype based on this markup have shown promising results . In the case of SPIRIT, further methods for ranking results also help to reduce the effects of incorrect markup.

REFERENCES

[1] GATE: A framework and graphical development environment for robust NLP tools and applications - Cunningham - 2002 (Show Context)

[2]Information extraction - Cowie, Lehnert - 1996 (Show Context)

[3]]Named Entity Recognition without Gazetteers - Mikheev, Moens, et al. - 1999 (Show Context)

[4]Web-a-where: geotagging web content - Amitay, Har'El, et al. - 2004 (Show Context)

[5]Computing Geographical Scopes of Web Resources - Ding, Gravano, et al. - 2000 (Show Context)

[6]Geospatial mapping and navigation of the web - McCurley - 2001 (Show Context)