Neogeography: The Challenge of Channelling Large and Ill-Behaved Data Streams

Mena B. Habib

Supervised by: Peter Apers and Maurice van Keulen
Database chair, University of Twente
Enschede, The Netherlands
m.b.habib@ewi.utwente.nl

***Abstract*—Neogeography is the combination of user generated data and experiences with mapping technologies. In this paper we propose a research project to extract valuable structured information with a geographic component from unstructured user generated text in wikis, forums, or SMSes. The project intends to help workers communities in developing countries to share their knowledge, providing a simple and cheap way to contribute and get benefit using the available communication technology.**

I. INTRODUCTION

Users are not passive recipients. Not only can they choose the type of information they want to access but also they can even produce the information themselves. The term "Neogeography", sometimes referred to as "volunteered geographic information (VGI)", is a special case of the more general web phenomenon of "user-generated content (UGC)", that has a relation to geographical features of the earth [1]. UGC refers to various kinds of media content, publicly available, that are produced by end-users. Such contents may include digital video, blogging, mobile phone photography, and wikis. UGC can provide citizens, consumers and students with information and knowledge as its contents tend to be collaborative and encourage sharing and joint production of information, ideas, opinions and knowledge among users.

In neogeography, end-users are not only beneficiaries but also contributors of geographic information. Neogeography combines the complex techniques of cartography and GIS and places them within reach of users and developers [2].

In this project, our wide objective is to propose a new portable, domain-independent XML-based technology that involves sets of free services that: enable end-users communities to express and share their spatial knowledge; extract specific spatial information; build a database from all the users' contributions; and make use of this collective knowledge to answer users' questions with sufficient level of quality. Users can use free text (such as SMS, blogs, and Wikis) to express both their knowledge and enquiries, or they can use map-assisted questions from their smart phones.

Users can benefit from this technology by using a question answering system to retrieve specific information about some place in the form of generated natural language and, if the communication device allows it, simple maps.

II. MOTIVATION

The rapid growth in the IT in the last two decades leads to the growth in the amount of information available on the World Wide Web. However, the information accessibility in the developing countries is still growing slowly. It is rare for a person in a developing country to have access to the internet. In Africa, which has a population of more than one billion, only one among every 250 persons has access to the internet [3]. On the other hand, figures released in March 2005 from the London Business School reported that Africa has seen faster growth in mobile telephone subscriptions than any other region of the world over the last five years. A recent study found 97 percent of people in Tanzania said they could access a mobile phone, while only 28 percent could access a landline [4].

The wide spreading of mobile phones coincides with developing applications and services based on wireless telecommunication. SMS text messaging can be an efficient and effective means of information sharing and accessing. The total number of SMS sent globally tripled between 2007 and 2010, from an estimated 1.8 trillion to a staggering 6.1 trillion [5].

The proposed system gives the workers' committees in developing countries, where governments are hardly covering the basic public services, the ability to help themselves, sharing their information through mobile phones. For example, truck drivers may provide the system with SMS messages about the traffic situation at particular places at a specific time. Structured information about the place, the time and the situation are extracted from these messages, and stored in a spatial DB. Users can benefit from this system by asking about the best way to go to somewhere by sending a SMS question.

Another possible application for this system concerns farmers' communities. Farmers can share their knowledge about climate changes, the suggested crops to be sown in a specific region or the possible markets into which they can sell their goods. Farmers can also keep track of the way a swarm of locusts is moving.

Many other applications in fields of health, urban utilities monitoring, and crisis management can be developed with our proposed system.

III. RELATED WORK

Within the previously discussed theme many projects have been developed to make use of users' contributions. Wikimapia

and OpenStreetMap are good examples of collaborative projects to create a free editable map of the world, while Google Earth and Flicker allow users to upload and place their own captured photos over the earth's map. Other tools like MapQuest and OpenAPI allow users to embed directions to some places in their web site. Users can share their directions, recorded by their GPS devices, using websites like GPSVisualizer and GeoTracing. Another application is "Digital Foot-printing" for tourists using the presence and movements from cell phone network data and the geo-referenced photos they generate [6]. Similarly, TwitterHitter plots the tweets of single twitter individual or group of individuals and generates an extended network graph view for visualizing connections among individuals in a region [7]. To bring this technology to the developing world, we need however to adapt it to the available communication technology, namely SMS on simple mobile phones.

Other research dealt with text as a source of geographic information. Numerous researches have focused on geo-parsing which tries to resolve geographic names appear in text [8][9]. "Places mentioned in this book" service provided by Google Books is one of those applications based on such researches. Other researches have tackled the area of analyzing and visualizing the frequencies of terms used in referring to geographical description [10][11]. Few researches try to model human natural language expression in representation of references to places [12][13]. Spatio-Temporal Information Extraction is mentioned by some researches for geographic information retrieval purposes [14][15][16]. The aim of those researches was to annotate documents with sets of locations and time information extracted from those documents, visualize this extracted information on digital map.

Other research groups worked on geographical ontologies. Within this paradigm, [17][18] focused on the problem of integrating multiple datasets for constructing geo-ontology for the purpose of developing a spatially-aware search engine, while [19] tried to propose a reference model for developing geographic ontologies. A GeoOntology Building Algorithm was developed by [20] to extract data from the different data sources (relational databases, XML documents, GML documents, etc.) and transform them into ontology instances. Similarly, [21] describes work done in order to integrate the information extracted from gazetteers, WordNet, and Wikipedia.

IV. Alternative Validation Scenario

Since there is no real-life service with historical data available to us at the beginning of this project, some alternative scenarios can be assumed for the project to prove its validity. Our suggested scenario is to use users' volunteered contents on Wikipedia to extract places names, places types, spatial relationships between geographical places, and thus to incrementally learn and instantiate a geographical ontology using this extracted information. According to [22] the quality of automatic knowledge acquisition is still significantly below that of a hand-crafted knowledge base. Our aim is to propose a high coverage with high quality geo-ontology. Resulting ontology can be compared to other existing geo-ontologies to evaluate its quality and coverage.

The relations between geographic places described in text have multiple forms. It can be either topological (ex: within, touches overlap, contains, etc.), directional (ex: east of, north west of, front of, etc.), or distance relation (ex: 5 km of, 30 min of, etc.). Mentioned places and relations are almost fuzzy, so ambiguity and vagueness can be expected to exist about the referred place. It is required to develop a framework which augments and supplements geo-ontology learnt with capabilities for representing and reasoning with uncertain and incomplete information.

The developed geo-ontology can be used in a tourism case study. Tourists are naturally motivated to share their experiences about a touristic destination, hotels' quality, transportation, prices, threats and so on, via forums like TripAdvisor. The system can extract useful information from these chunks of texts and formulate it in a structured way. This information can be the users' suggestions for popular places that are worth visiting in some city. The extraction process can make use of the geo-ontology generated from wikipedia as part of the interpreting process for extracting the relevant information

After the extraction process, the extracted information should be integrated into a probabilistic DB using a probabilistic framework to deal with the uncertainty that comes with the users' contributions. As we are dealing with opinions expressed by actual humans, much contradiction and subjective uncertainty can be expected, which requires that the entire process must support handling of probabilistic data.

The system users can benefit from this data by submitting queries like "Touristic places within Paris" using question answering mechanism. The system then will use the extracted information with the help of the generated geo-ontology to answer those questions.

V. Challenges and Research Questions

This project is at the cross roads of several research areas. These research areas include: information extraction, the semantic web, probabilistic data integration, probabilistic XML databases, and spatial databases.

Information Extraction (IE) plays a major role in this project. IE systems analyse human language text in order to extract information about pre-specified types of events, entities or relationships. In our case, the users' community keeps providing their knowledge about conditions within particular geographic regions in a dynamic, free-text manner and our task is to extract valuable information from this mass of text and use it to populate a pre-specified templates. This requires the extraction of the W4 questions of: *who*, *where*, *when* and *what* from textual descriptions.

Information Extraction from text sources is, by nature, challenging in many ways:

- Information contained in text is often partial, subject to evolution over time, in conflict with other sources, and sometimes untrustworthy.
- Recognizing the coreference of entities and events when they are described or referred to in different textual sources.

- The lack of clear guidelines for evaluating the correctness of the output generated by an extraction algorithm.

The nature of the project implies some other challenges:

- Information about spatial data adds another challenge of resolving the spatial vagueness. Some places have the same names and sometimes the spatial information is not well defined, or changes from time to time.
- Different textual sources imply different ways of writing, and expression.
- IE systems are always built for a specific domain. Research is required on the automatic acquisition of template filler patterns, which will enable systems for much larger domains.

Uncertainty in data is another challenge point. Uncertainty may come in different ways:

- Uncertainty in the extraction process, i.e. the precision level expected from the IE system in resolving facts or geographical names.
- Uncertainty in the source of information, i.e. the possibility that the data provided is completely or partially incorrect.
- The contradictions between the extracted information and the information previously extracted and stored in the probabilistic database.
- The validation of the information over time. Geographical information is dynamic information and always changing over time.

Semantic web and linked data must have precedence when we are dealing with global neogeographic systems. The semantic web adds another challenge of linking the rapidly growing number of existing web data sources to find the meaningful content [23].

There is a growing interest in designing probabilistic XML-databases to represent uncertain data [24]. Besides, spatial databases support spatial data types in its implementation, providing spatial indexing and spatial join methods. In this project, it is strongly needed to make use of both mentioned types of databases by extending the probabilistic XML-databases with capabilities to represent spatial information.

Solving these problems calls for ideas from multiple disciplines, such as machine learning, natural language understanding, machine translation, probabilistic data integration, knowledge representation, data management, and linguistic theory related to language semantics.

Based on challenges stated above, our research project will try to find answers to the following research questions:

- How much coverage and data quality can be achieved for a geo-ontology learned from textual content compared to manual geo-ontologies?
- How to extend geo-ontology with capabilities for representing and reasoning with uncertainty?
- How can more domain portability for IE technology be realized?
- What probabilistic framework can manage uncertainty in the IE process?

VI. Proposed System Architecture

Figure 1 shows the proposed system architecture for the project.

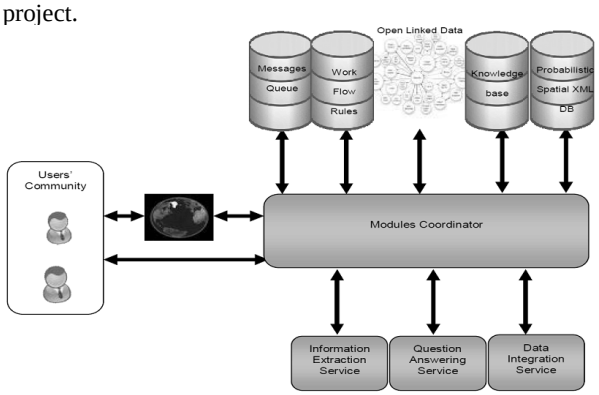

Fig. 1. The proposed system architecture.

1) Modules Coordinator (MC)

This module is the controller of the whole system. It is responsible for controlling the work and data flow between different services. It receives the user contributions and requests, and sends activation messages to the intended services according to set of workflow rules.

2) Information Extraction (IE) Service

This is the key service of the system. This module reads input text from the messages queue, checks if the message contains information or a question, and in response sends the type of the message to the MC to determine the suitable workflow. In both cases, the IE is responsible for processing the text message. If it is an information message, the IE service reads the extraction rules from the Knowledge base, tries to extract the required information from the textual data, assigns some certainty factor to the extracted information and then passes this extracted information to the data integration service. In the case of a request message, the IE service then has to understand what this question wants to find and passes the request keywords to the question answering system.

3) Data Integration (DI) Service

Data integration task comes in two ways. The first is to integrate the information provided by the IE service with the information already existing in the XML Database (XMLDB). It tries to find the information in the XMLDB that refers to the same geographical place mentioned by the IE, finds the conflicting facts, and tries to resolve such conflicts using the knowledgebase (KB) independently of the user by assigning several levels of certainty to each new piece of information. The second is to manage integrating data from Open Linked Data (OLD) web ontologies in a consistent and efficient manner to achieve the goals of the project. Data integration over OLD also implies uncertainty in the integrated data.

4) Question Answering (QA) Service

This service receives the request keywords from the IE service, formulates the XML query, runs this query on the DB, retrieves the results, applies some inference on the results using geo-ontology if needed and sends the results back to the user in the form of natural language generated text.

5) Probabilistic Spatial XML Database (XMLDB)

This database is a standard probabilistic XML DB that is extended to handle geospatial data. The information contained in this DB is assigned to some certainty factor that indicates how certain the information is. The data integration module is responsible for assigning this certainty factor.

6) Knowledge Base (KB)

Holds set of rules needed for the extraction process. These rules are generated from a set of training texts. Also, it handles the probabilistic framework used for assigning probabilities to the possible locations, resolving conflicts between extracted information and those existing in the XMLDB.

7) Open Linked Data (OLD)

All the modules make use of web ontologies to enrich and improve the data.

8) Message Queue (MQ)

The queue of text messages received from users that need to be processed.

9) Work Flow Rules (WFR)

These are the rules for activating intended modules on the basis of the type of message being processed.

VII. CONCLUSION AND FUTURE WORK

This paper proposes a system to manage the collective knowledge that resides in users' community on a particular domain. Users can share their knowledge, and ask for information in free text. The system extracts structured information out of users' contributions, and makes it available for users upon request. The project has to overcome set of challenges comes with IE by providing a probabilistic framework to handle uncertainty in the extracted information. Spatial capabilities must be added to probabilistic XML-databases to deal with vagueness geographic places described in text.

In the near future work we are going to investigate possible solutions to the presented research questions. A prototype is to be developed and published as a free web service to prove the validity and the visibility of the system. In the far future we want to apply the system on a set of real-life domains.

REFERENCES

[1] M. F. Goodchild. "Citizens as sensors: the world of volunteered geography". GeoJournal, Vol. 69, No. 4. Pages 211-221. 2007.

[2] Andrew J. Turner. "Introduction To Neogeography". O'Reilly Media, Inc. 2006.

[3] Mike Jensen. "The outlook for the telecentres and cybercafes in Africa". http://www.acacia.org.za/jensen_articles.htm.

[4] Rhett Butler. "Cell phones may help "save" Africa". http://news.mongabay.com/2005/0712-rhett_butler.html.

[5] "The World in 2010: ICT facts and figure". International Telecommunication Union. 2010.

[6] F. Calabrese, F. D. Fiore, C. Ratti, J. Blat. "Digital Footprinting: Uncovering Tourists with User-Generated Content". IEEE In Pervasive Computing, Vol. 7, No. 4. Pages 36-43. 2008.

[7] Jeremy J. D. White, Robert E. Roth. "TwitterHitter: Geovisual analytics for harvesting insight from volunteered geographic information". In Proceedings of GIScience 2010. 2010.

[8] Bertrand De Longueville, Nicole Ostländer, and Carina Keskitalo. "Addressing vagueness in Volunteered Geographic Information (VGI) – A case study". International Journal of Spatial Data Infrastructures Research, Vol 5. 2010.

[9] Simon E Overell. "Geographic Information Retrieval: Classification, Disambiguation and Modelling". PhD thesis. University of London. 2009.

[10] J. Dykes, R. Purves, A. Edwardes, and J. Wood. "Exploring Volunteered Geographic Information to Describe Place: Visualization of the 'Geograph British Isles' Collection". In Proceedings of the GIS Research UK 16th Annual Conference GISRUK 2008. Pages 256-267. 2008.

[11] Ross Purves, Jason Dykes, Alistair Edwardes, Livia Hollenstein, David Mueller and Jo Wood. "Describing the space and place of digital cities through volunteered geographic information". GeoVis workshop in Hamburg-Germany. 2009.

[12] I. Mani, J. Hitzeman and C. Clark. "Annotating natural language geographic references". In proceeding of LREC 2008-W13 Workshop on Methodologies and Resources for Processing Spatial Language. Pages 11-15. 2008.

[13] C. Sallaberry, M. Gaio, J. Lesbegueries and P. Loustau. "A Semantic Approach for Geospatial Information Extraction from Unstructured Documents". The Geospatial Web, Advanced Information and Knowledge Processing. Springer. Pages: 93-104. 2007.

[14] Y. Chen, G. Di Fabbrizio, D. Gibbon, R. Jana, and S. Jora. "GeoTracker: Geospatial and Temporal RSS Navigation". In Proceedings of the 16th international conference on World Wide Web. Pages: 41-50. 2007.

[15] B. Martins, H. Manguinhas, and J. Borbinha. "Extracting and Exploring the Geo-Temporal Semantics of Textual Resources". In Proceedings of the IEEE International Conference on Semantic Computing. Pages 1-9. 2008.

[16] J. Strötgen, and M. Gertz. "TimeTrails A System for Exploring SpatioTemporal Information in Documents". In Proceedings of the 36th International Conference on Very Large Data Bases. 2010.

[17] G. Fu, C. B. Jones and A. I. Abdelmoty. "Building a Geographical Ontology for Intelligent Spatial Search on the Web". In Proceedings of IASTED International Conference on Databases and Applications. Pages 167-172. 2005.

[18] F. J. Lopez-Pellicer, M. Chaves, C. Rodrigues, and M. J. Silva. "Geographic Ontologies Production in GREASE-II," University of Lisbon, Faculty of Sciences, LaSIGE, Tech. Rep. TR 09-18. 2009.

[19] G. N. Hess, C. Iochpe, and S. Castano. "Towards a Geographic Ontology Reference Model for Matching Purposes". In the proceedings of the 9th Brazilian Symposium on GeoInformatics. Pages 35-47. 2007.

[20] Y. Wei, C. Jiaheng, L. Qing, and C. Junpeng "Ontology-based Geographic Information Retrieval and Ranking". In the International Semantic Web Conference ISWC'06 Workshop. 2006.

[21] D. Buscaldi , P. Rosso , and P. P. García. "Inferring geographic ontologies from multiple resources for geographic information retrieval". In SIGIR Workshop on Geographic Information Retrieval. Pages 52-55. 2006.

[22] F. Suchanek, G. Kasneci, and G. Weikum. "YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia". In Proceeding of World Wide Web Conference. Pages 697–706. 2007.

[23] V. Richard Benjamins and Jesús Contreras and Oscar Corcho and Asunción Gómez-pérez. "Six Challenges for the Semantic Web". In KR2002 Semantic Web Workshop. 2002.

[24] T. Li, Q. Shao, and Y. Chen. "PEPX: a query-friendly probabilistic XML database". In Proceedings of the 15th ACM international conference on Information and knowledge management. Pages: 848 – 849. 2006.