UNIVERSITY OF TWENTE.

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

Technical Report

Authors: Mena B. Habib
Dr.ir. Maurice van Keulen

Faculty \ Department: EWI \ Database

Abstract

Neogeography is the combination of user generated data and experiences with mapping technologies. In this article we present a research project to extract valuable structured information with a geographic component from unstructured user generated text in wikis, forums, or SMSes. The extracted information should be integrated together to form a collective knowledge about certain domain. This structured information can be used further to help users from the same domain who want to get information using simple question answering system. 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.

Table of Contents

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].

Within this theme many projects have been developed to make use of users' contributions. Wikimapia[1] and OpenStreetMap[2] are good examples of collaborative projects to create a free editable map of the world, while Google Earth[3] and Flicker[4] allow users to upload and place their own captured photos over the earth's map. Other tools like MapQuest OpenAPI[5] 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[6] and GeoTracing[7]. 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 [3]. 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 [4]. 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 [5][6][7]. "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 [8][9]. Few researches try to model human natural language expression in representation of references to places [10][11]. Spatio-Temporal

[1] http://wikimapia.org
[2] www.openstreetmap.org
[3] earth.google.com
[4] www.flickr.com
[5] www.mapquest.com
[6] www.gpsvisualizer.com
[7] www.geotracing.com

Information Extraction is mentioned by some researches for geographic information retrieval purposes [12][13][14][15]. 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, [16][17] focused on the problem of integrating multiple datasets for constructing geo-ontology for the purpose of developing a spatially-aware search engine, while [18] tried to propose a reference model for developing geographic ontologies. A GeoOntology Building Algorithm was developed by [19] to extract data from the different data sources (relational databases, XML documents, GML documents, etc.) and transform them into ontology instances. Similarly, [20] describes work done in order to integrate the information extracted from gazetteers, WordNet, and Wikipedia.

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 a 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.

This technology involves extracting information from semi-structured or even unstructured text with help of Web2.0 capabilities and transferring this information to a probabilistic XML database which manages the uncertainty in the extracted information. The proposed technology has to be portable and domain-independent so that only minor changes need to take place before applying this technology to each new scenario.

The main core of the project involves extracting information from semi or unstructured text. Unfortunately, Information Extraction (IE) from textual sources is challenging in many ways. Information contained in text is often partial, subject to evolution overtime, in conflict with other sources, and sometimes untrustworthy. Also, resolving spatial vagueness in geographic data is another issue. Finally, IE systems are always built for a specific domain, so it is challenging to have a portable IE system that can be applied on different domains with only minor customization.

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 [21].

Since computers are rare in much of the developing countries due to poor wire-line infrastructure, 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 [22].

The wide spreading of mobile phones coincides with developing applications and services based on wireless telecommunication. SMS text messaging, the most widely used data application with 5.3 billion active users [23], 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. In other words, close to 200 000 text messages are sent every second [23].

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, linked with some other GISs and then 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 plants' blights or 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.

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. Any suggested scenario must satisfy three requirements. First, it should deal with user generated data that contains geospatial information. Second, data should be short text to imitate SMS messages. Finally, data should contribute to one domain, and aggregates to collective knowledge.

Our selected scenario is the tourism scenario. Tourists are naturally motivated to share their experiences about a touristic destination, hotels' quality, transportation, prices, threats and so on, via forums, blogs or even twitter[8] messages. Twitter messages will be used as the source of information because it satisfies all requirements mentioned for the validation scenario. The system can extract useful information from these tweets and represent it in a structured way. This information can be some basic information about the hotel like its classification or its average room price per night or even be the users' opinions about hotels services. The system should extract the hotel name, infer and disambiguate its location, and should extract also the piece of information associated with this hotel. The extraction process can make use of the existing geo-ontologies 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 informal text containing opinions expressed by actual humans, much contradiction and subjective uncertainty can be expected, which requires that the entire process should support handling of probabilistic data.

The system users can benefit from this data by submitting queries like "What are the good/cheap hotels near Paris?" using question answering mechanism. The system then will use the extracted information with the help of geo-ontologies to answer those questions.

[8] http://twitter.com

Problem Statement, Challenges and Research Questions

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

Information Extraction 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 co-reference 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 [24].

The nature of the project implies some other challenges:

- As the type of information\event stated in the volunteered user text is unexpected, it is hard to extract specific information. In other words, we don't know exactly what kind of information is stated in the text and hence try to extract it out. Current IE approaches try to extract specific predefined events, but in our case extra efforts is required to infer the type of information expressed in the text.
- 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. For example "Cairo" is the name of more than ten cities and other geographic places such as lakes and mountains.
- The short length of some texts, such as SMS messages makes the inferring process harder. Language used in short messages is always informal and has many modern new abbreviations and expressions and sometimes contains misspelling.
- Different textual sources imply different ways of writing, and expression.
- IE systems are always built for a specific domain, so that it is challenging to have a portable IE system that can be applied on different domains with only minor customization. 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. Making use of these sources simultaneously is beneficial and challenging. In one way, it extends the current human-readable web by encoding some of the semantics of resources in a machine-processable form. Making computers able to search, process, integrate and present the content of these resources in a meaningful, intelligent manner. On the other hand, we need to worry about how to manage all this data in a scalable manner, that is, how to find the meaningful content among this huge mass of available information [25].

There is a growing interest in designing probabilistic XML-databases to represent uncertain data [26]. 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, 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:

Q1. Could the existing IE techniques be applied successfully to short informal abstract messages? How can IE techniques be adapted or extended to suit short messages?

Q2.a. Will the natural language processing techniques (POS tagger, Syntactic analyzer, .. etc.) perform as adequate as they should on informal text?

Q2.b. What features can be used for Named Entities (names of persons, organizations, locations, expressions of times, quantities, etc) extraction in informal short text?

Q2.c. What methods can be used for Named Entities disambiguation in informal short text?

Q2.d. How to infer about the referred location from relative references (like: "north of", "in vicinity of")?

- **Discussion:** Named entity extraction is subtopic of IE which aims to extract entity names (for people and organizations), geographic names, temporal expressions, and certain types of numerical expressions. Multiple approaches were used in traditional IE to extract named entities. Most of those approaches used Gazetteers Lookup, Shallow Parsing, which uses internal (like part of speech tags and Orthographic features) and external evidence (like the local context), to extract the named entity. However, in short informal text, such approaches are not expected to perform well due to the informality nature of the text. For example: the tweet "*obama should b told NO vote on tax deal unless omnibus is made public in advance* !", lacks the use of capitalization for the proper noun "*Obama*", also the abbreviation "*b*" is used instead of "*be*". Although solving this abbreviations problem is out of our research focus, it affects the state of the art of IE which uses part of speech tagging. Upon this problem, new features should be discovered to enable named entity extraction.

  Consider also the tweet "*Fox Sports Grill is a few blocks north of your hotel, Lola is next to the restaurant, McCormick & Schmicks is a few blocks west*". It is required to investigate how to guess, with high level of certainty, the hotel mentioned in the text using the clews "*Fox Sports Grill is a few blocks north of your hotel*" and "*McCormick & Schmicks is a few blocks west*"?

  Once named entity is extracted, another problem comes to surface. This problem is the disambiguation of the named entity. Here we will take geographic names as an example. Geographic names, like all other Named Entity, are highly ambiguous. For example, the famous capital "Paris" refers to 62 different geographic places around the world, "San Antonio" refers to 1561 references. Table 1 shows the top ten of the most ambiguous geographic names in geonames database[9]. Also figures 1 and 2 show the long tail distribution of the ambiguous geonames and the percentage of geographic names with multiple references in geoname database respectively.

Table 1: The most ambiguous geographic names in geonames database.

Geographic name	Number of references
First Baptist Church	2382
The Church of Jesus Christ of Latter Day Saints	1893
San Antonio	1561
Church of Christ	1558
Mill Creek	1530
Spring Creek	1486
San José	1366
Dry Creek	1271
First Presbyterian Church	1229
Santa Rosa	1205

[9] www.geonames.org/

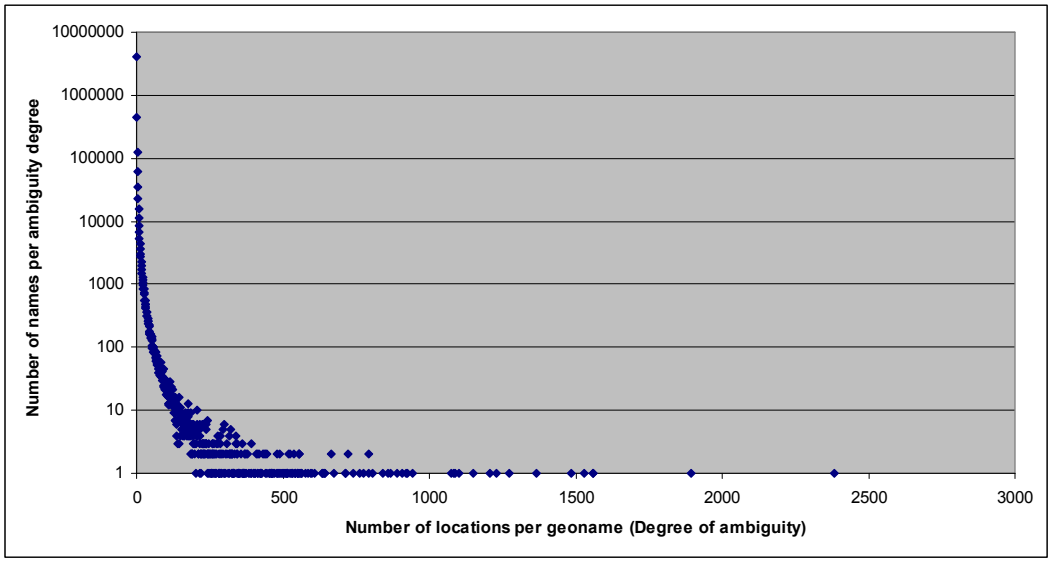

Figure 1: Ambiguity of geographic names in geonames database

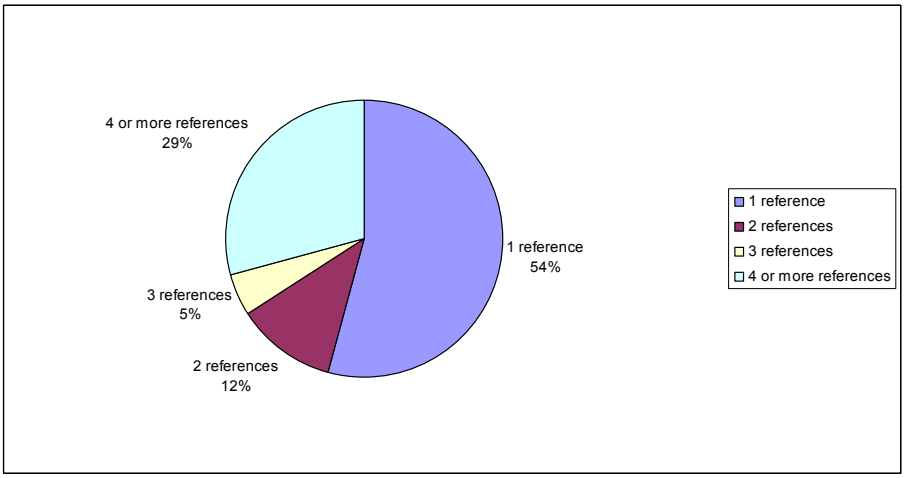

Figure 2: Percentage of geographic names with multiple references in geoname database

Another challenge of geographic names disambiguation is the inference of the geographic place using references. 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 relations are always fuzzy and comprise vagueness. For example terms like “nearby”, “north of”, or “in vicinity of”, are vague and implies some degree of uncertainty about the referred place. All those challenges dealing with named entity extraction and resolution require development of a framework which augments and supplements

the extraction system with capabilities for representing and reasoning with uncertain and incomplete information.

To answer this set of research questions it is required to study the capability of traditional techniques of named entity extraction to suit the informal text. We believe that some named entities (like numeric quantities and dates) can still be extracted with traditional techniques; however we still need to evaluate this capability to all other named entities to find to what extent it is suitable to each type of named entities.

As there will be some named entities can not be extracted by traditional techniques, we have to find the reasons that hamper the extraction. We need also to find new features that could help on extraction of those named entities. As an extension to this problem, comes the problem of disambiguation. All disambiguation approaches uses the context as a guide. However, short text is lack of enough context and hence we to find another clues for disambiguation. The solutions for both of the extraction and disambiguation should look for new features that may be related to linguistics, semantics or users profiles.

Q2. What probabilistic framework can manage uncertainty in the IE/DI process?

- Q2.a. What are the sources of uncertainty in the extracted information?
- Q2.b. How to measure different sources of uncertainty?
- Q2.c. How to combine those measures to model the uncertainty of the whole process?
- Q2.d. How to make use of the combined uncertainty measures to improve integration of extracted information with those already exist in the database?
- Q2.e. How to make use of the existing Open Linked Data and web resources in a specific domain to improve extracted information certainty?
- Q2.f. How dealing with uncertainty can improve both the extraction and disambiguation processes?

- **Discussion:** As discussed before, uncertainty can arise for multiple reasons. For information extraction, current techniques are often based on statistical machine learning methods and can be extended to compute probabilities of each extraction result. If we have tweets like this: "*Essex House Hotel and Suites from $154 USD*" and "*Essex House Hotel and Suites from $123 USD: Surrounded by clubs and designer*". First, we are uncertain about the hotel name itself, either it is "*Essex House Hotel*" or "*Essex House Hotel and Suites*". Second, we should resolve the hotel entity name. According to tripadvisor[10], 5 hotels are named "*Essex House Hotel*" in different locations in the United States. Another cause of uncertainty is the contradicting fact of the minimum price among the two messages. There may be also some uncertainty about how trustful are the users

[10] www.tripadvisor.com/

who sent those messages. Another type of uncertain data is the reference to geo-locations. As discussed in the first research questions, the problem of resolving geographic names is very ambiguous.

To answer this set of research questions we have to identify the sources of uncertainty, and to measure their impact on the extracted data quality individually (one by one) to specify the different weights for those sources. These measures of uncertainty should be combined together to formulate the final uncertainty level of the extracted and the integrated information. The existing web sources of information like ontologies, gazetteers and encyclopedias can be used to enhance the certainty levels of the disambiguation process which can be used further to enhance the extraction process in an iterative manner.

According to the literature review done in [27] it was clear that those researches done in fields of information extraction; data integration and uncertainty management are still stand alone. Information extraction researches focus on how to improve the extraction and the disambiguation process, or how to generalize the process unsurprisingly across multiple domains. Researches on data integration focus on methods to enhance schema matching; duplicate detection and data fusion. On a third direction, Uncertainty management deals with only individual kinds of uncertainty. However, the discussed challenges and research questions shows the need of combining those fields of research. Dealing with different sources of uncertainty can be a key factor to enhance the information extraction, disambiguation and integration.

Proposed System Architecture

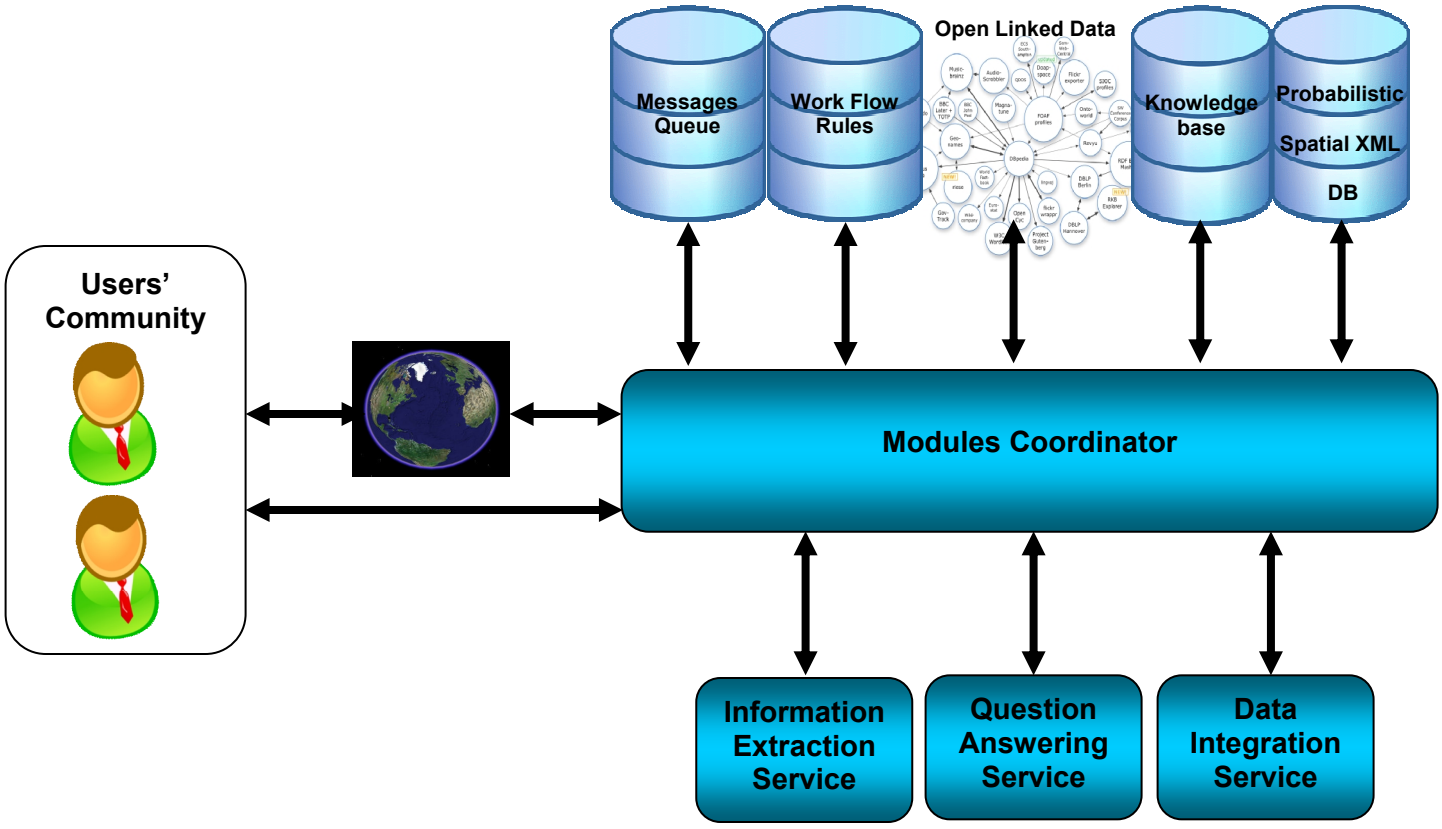

Figure 3: Proposed System Architecture

I. Modules description

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.

Information Extraction (IE) Service: This is the key service of the system. This module reads its input text from the messages queue and 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.

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.

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.

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.

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.

Open Linked Data (OLD): All the modules make use of web ontologies to enrich and improve the data.

Message Queue (MQ): The queue of text messages received from users that need to be processed.

Work Flow Rules (WFR): These are the rules for activating intended modules on the basis of the type of message being processed.

- In the case of a received message, the MC passes a signal to the information extraction module that there is a new message in the message queue that needs to be processed. The information extraction (IE) module processes the message, identifies its type (information or request) and sends a signal back to the MC defining the type of the message.
- In the case of receiving a message containing information, the IE module processes the message and extracts implicit information with the help of the open linked data. The extracted information is then passed then to the data integration module which makes use of the knowledgebase and the existing data in the XML database to integrate the extracted information into the XML probabilistic DB.
- In the case of a request sent by the user, the IE module extracts the key words from the request and then uses these key words to generate a formal XML query, runs this query over the DB and finally forwards the results to the MC to be sent to the user.

II. Example of a possible scenario

Here we present an example of one possible scenario to see how the system will run. In this demonstration we deal with the tourism domain. Users send the following messages to our system:

"*berlin has some nice hotels i just loved the hetero friendly love that word Axel Hotel in Berlin.*"

"*Good morning Berlin. The sun is out!!!! Very impressed by the customer service at #movenpick hotel in berlin. Well done guys!*"

"*In Berlin hotel room, nice enough, weather grim however*"

Once a message is received, it is placed in the **MQ**. A signal is sent to the **MC** indicating that there is a new message is waiting for processing. The **MC** will then activate the **IE** module which fetches the message from the **MQ**, and classifies it message as an "Informative Message". A tag is then attached to the message on the **MQ** indicating its type. The **MC** checks the set of **WFR** according to the type of the message. The **IE** module is activated again by the **MC** to extract the information that is implied in the message. The **IE** uses the extraction rules stored in the **KB** to extract the required information with some degree of certainty about the city and country names using Geo-ontology and with the aid of other geographical names contained within the message.

Let us say that we want to extract users' attitude against some hotel along with its location (city). In this case, the IE will extract the following templates:

Field	Template 1	Template 2	Template 3
Hotel_Name	*Axel Hotel*	*movenpick hotel*	*Berlin hotel*
Location	*Berlin*	*Berlin*	*Berlin*
Country	*P(Germany)>*	*P(Germany)>*	*P(Germany)>*
	P(USA)>	*P(USA)>*	*P(USA)>*
	P(…)	*P(…)*	*P(…)*
User_Attitude	*P(Positive)>*	*P(Positive)>*	*P(Positive)>*
	P(Negative)	*P(Negative)*	*P(Negative)*

A signal is sent back from the **IE** module to the **MC** signalling the end of the extraction process. The **MC** activates the **DI** module which receives the extracted information from the IE module. The **DI** module then is responsible of resolving the conflicts between the extracted information and the existing information. In the case that same information already exists in the **XMLDB**, the **DI** module has to modify the certainty factor attached with the existing information using set of rules in the **KB**. In the case that the extracted piece of information is totally novel information, the **DI** module adds a record to the *Hotels* table in the **XMLDB** and attach a certainty factor based on the precision of the extraction of the locations names.

Now let us examine the other scenario, that of a user request. The user sends the following request:

"*Can anyone recommend a good, but not ridiculously expensive hotel right in the middle of Berlin?*"

The message is placed in the **MQ** and the **IE** is activated to indicate the type of the new message. The **IE** module marks the message as a "Request Message". The **WFR** associated with "Request Message" is used by the **MC**. The **IE** extracts the keywords of the request (*hotel, Berlin, good, not expensive).* Then the **QA** module is activated and receives those keywords and formulates the suitable XQuery (assuming the existence of functions like *topk, score*):

```
topk(3, for $x in // Hotels
where $x/City == "Berlin" and $x/ User_Attitude == "Positive"
orderby score($x)
return $x)
```

The XQuery is applied on the **XMLDB** and the retrieved records are passed back again to the **QA** module which uses those records to form a natural language answer to the users request. The answer is forwarded to the **MC,** which in turn forwards it to the user.
The answer may be like this:

"*Some good hotels in Berlin are Axel Hotel, movenpick hotel, Berlin hotel.*"

Research Plan

Year	From	To	Tasks
First	May 2010	October 2010	• Research project proposal. • Information Extraction (IE) literature review. • Research on IE over GATE tool. • Publishing one page proposal on DBDBD.
	November 2010	January 2011	• Publishing project proposal as a short paper in a conference or in a PhD workshop. • Probabilistic Data Integration literature review. • Modelling uncertainty in Information Extraction literature review. • Research on Geographic Names Disambiguation.
	February 2011	April 2011	• Publishing paper about "Geographic Names Disambiguation from Uncertainty Prospective". • Data collection and filtering. • Applying IE techniques on twitter messages.
Second	May 2011	April 2012	• Research to find answer to research questions 1 and 2. • Publishing at least two papers about the developed solutions. • Attend summer school.
Third	May 2012	April 2013	• Research on extended directions. • Publishing at least two papers about the developed solutions. • Research visit to other research group.
Fourth	May 2013	April 2014	• Integrating the system contributions. • Publishing a collective journal paper. • Thesis writing. • Defence.

References

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

[2] A. J. Turner. “Introduction To Neogeography”. O'Reilly Media, Inc. 2006.

[3] 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.

[4] J. J. D. White, R. E. Roth. “TwitterHitter: Geovisual analytics for harvesting insight from volunteered geographic information”. In Proceedings of GIScience 2010. 2010.

[5] B. De Longueville, N. Ostländer, and C. Keskitalo. “Addressing vagueness in Volunteered Geographic Information (VGI) – A case study”. International Journal of Spatial Data Infrastructures Research, Vol 5. 2010.

[6] S. E. Overell. “Geographic Information Retrieval: Classification, Disambiguation and Modelling”. PhD thesis. University of London. 2009.

[7] J. L. Leidner. “Toponym Resolution in Text. Annotation, Evaluation and Applications of Spatial Grounding of Place Names”. PhD thesis. University of Edinburgh. 2007.

[8] 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.

[9] R. Purves, J. Dykes, A. Edwardes, L. Hollenstein, D. Mueller and J. Wood. “Describing the space and place of digital cities through volunteered geographic information”. GeoVis workshop in Hamburg-Germany. 2009.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] J. Strötgen, M. Gertz, and P. Popov. “Extraction and exploration of spatio-temporal information in documents”. In Proceedings of the 6th Workshop on Geographic Information Retrieval. 2010.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] M. Jensen. “The outlook for the telecentres and cybercafes in Africa”. http://www.acacia.org.za/jensen_articles.htm.

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

[23] “The World in 2010: ICT facts and figure”. International Telecommunication Union. 2010.

[24] A. De Sitter, T. Calders, and W. Daelemans. “A Formal Framework for Evaluation of Information Extraction”. Technical Report, University of Antwerp, Dept. of Mathematics and Computer Science. 2004.

[25] 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.

[26] 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.

[27] Mena B. Habib and M. van Keulen. “Information Extraction, Data Integration, and Uncertain Data Management: The State of The Art”. Technical Report TR-CTIT-11-06, Centre for Telematics and Information Technology, University of Twente. 2011.