Geographic Knowledge Discovery from Web 2.0 Technologies for Advance Collective Intelligence

Ickjai Lee and Christophcer Torpelund-Bruin
School of Business (IT), Cairns Campus, QLD 4870, Australia
E-mail: {Ickjai.Lee, Christopher.Torpelund}@jcu.edu.au

Keywords: geographic knowledge discovery, collective intelligence, Web 2.0 technologies, data mining, decision support

Received: February 24, 2011

Collective intelligence is currently a hot topic within the Web and Geoinformatics communities. Research into ways of producing advances with collective intelligence is becoming increasingly popular. This article introduces a novel approach to collective intelligence with the use of geographic knowledge discovery to determine spatially referenced patterns and models from the Geospatial Web which are used for supporting decisions. The article details the latest Web 2.0 technologies which make geographic knowledge discovery from the Geospatial Web possible to produce advanced collective intelligence. The process is explored and illustrated in detail, and use cases demonstrate the potential usefulness. Finally, potential pitfalls are discussed.

Povzetek: Članek se ukvarja z obdelavo geografskih podatkov s tehnologijo spleta 2.0.

1 Introduction

The second generation of web development and design have been coined as Web 2.0 technologies, where 2.0 refers to the historical context of web businesses coming back after the dot-com collapse [16]. Web 2.0 incorporates the move from Web-as-information-source architecture to the concept of Web-as-participation-platform, whereby users are encouraged to add value to the application as they create and collaborate information. The openness and freedom of user participation paves the way for Collective Intelligence (CI) which allows applications to be continuously improved to deeper the relationship with the users. This cycle of improvement is known as the perpetual beta, where a final version of the application is never reached - it simply continues to become better by offering more targeted experiences for each user according to their personal need [2].

As users of Web 2.0 services have grown, the functionality that the services provide have evolved into real world oriented human functions [17, 19, 22]. This implies the merging of geographical information with the abstract information that currently dominates the Internet. Often is the case that a user will search for something based on added spatial and temporal constraints. For example, "what is the best restaurant closest to a location x?", or, "how long will it take to get to the nearest hospital?". The merging of information with the real world has been dubbed as the Geospatial Web or Geoweb for short. The current explosion of digital geographic and geo-referenced datasets is said to be the most dramatic shift in the information environment for geographic research since the Age of Discovery [14]. Virtual globes such as Google Earth and NASA World Wind as well as mapping websites such as Google Maps, Live Search Maps and Yahoo Maps have been major factors in raising awareness towards the importance of geography and location as a means to index information. However, current collective intelligence techniques often fail to take into account these added spatial and temporal dimensions on user interactions and contributions. By considering these added dimensions, particular patterns and knowledge could be discovered about the users which could improve the accuracy of collective intelligence techniques.

Producing collective intelligence is a difficult challenge with the already vast amounts of user generated datasets on the Internet. The problem can become complicated when dealing with datasets with added geographic dimensions. The following are potential challenges associated with Geographic Knowledge Discovery (GKD) on spatially referenced datasets: 1) data access (inaccessibility) challenge; 2) diverse data types (inconsistency) challenge; 3) user interface (unavailablity) challenge. What has been proposed to overcome some of the challenges related to GKD is the need for a solid geographic foundation that accommodates the unique characteristics and challenges presented by geospatial data. Current national and global geospatial data lacks a proper infrastructure whereby contributed data can be aggregated and fully utilized for CI.

We propose to explore the use of GKD as a new technique for generating CI. GKD is an extension of Knowledge Discovery from Databases (KDD) and is based on a belief that there is novel and useful geographic knowledge hidden in the unprecedented amount and scope of digital geo-referenced data being collected, archived and shared by researchers, public agencies and the private sector [14].

Previous work [20] briefly explores the GKD model from Geoweb whilst current work extends it to extensive collective intelligence through GKD from Geoweb. Using the Voronoi diagram for Geoweb for emergency management [21] has been reported and algorithmic aspect of web map segmentation has been reported [11]. In this article, we investigate new emerging Web 2.0 technologies and whether they provide a means for generating the foundation that can be used to conduct GKD effectively to produce highly accurate CI for profitable traffic. The main aim of this article is not to empirically evaluate the performance of proposed framework with recommender systems and other visualization approaches, but to illustrate how Web 2.0 and Geoweb technologies could be used for GKD processes and advanced CI. Our proposed advanced CI process is abstractly described in Figure 1. Web 2.0 technologies are particularly used for user-oriented data selection and visualization. The proposed CI process can be used as an exploratory tool rather than a confirmatory tool. The main aim of this article is not to quantitatively compare and contrast with recommender systems, but to illustrate GKD processes from Web 2.0 and Geoweb technologies for advanced collective intelligence. Case studies show that how the generalized Voronoi diagrams and clustering can be combined and used with user-oriented datasets available through Web 2.0 and Geoweb. They demonstrate the potential usefulness and applicability of our proposed framework. The structure of the remainder of this article is as follows: Section 2 provides an overview of the latest Web 2.0 technologies which can be used to integrate the process. Section 3 describes the KDD processes along with the GKD for CI process. Section 4 demonstrates case studies using the GKD for CI process. Finally, section 5 concludes the article by listing potential pitfalls of the proposed process.

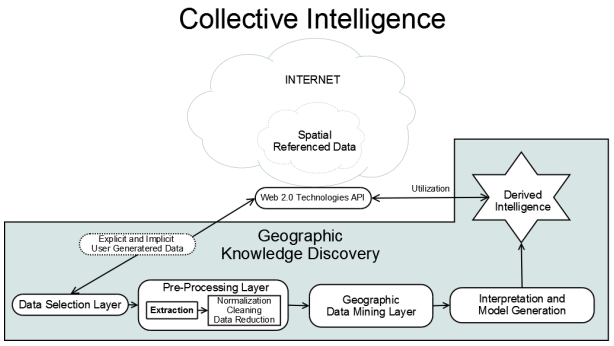


Figure 1: GDM on Web 2.0 technologies for CI process structure.

2 Web 2.0 Technologies

Web 2.0 technologies cannot be summed up and generalized but are instead a complex and continually evolving technology infrastructure which can include server-software, content-syndication, messaging-protocols, standards oriented browsers with plug-ins and extensions, and various client-applications. The encapsulating services may use just one or a combination of technologies, as the models defining the technologies are designed for hackability and remixability following open standards [2]. This necessitates fewer restrictions and allows for wider adoption and reuse. This infrastructure of complementary technologies provide services with information-storage, creation, and dissemination challenges and capabilities that go beyond what the public formerly expected in the environment of the so-called "Web 1.0". With the capabilities come the possibilities CI, but only if the challenges are overcome by the wide spread adoption of open Web 2.0 standards. Some of the common and standard Web 2.0 technologies used in the paper include:

- **Folksonomies:** The ability to allow collaborative tagging, social classification, social indexing, and social tagging.

- **Extensible Markup Language (XML) and/or Java Script Object Notation (JSON):** A general purpose specification for creating custom markup languages. Its primary purpose is to help share structured data based on user defined elements. XML document are compiled with a particular schema/Data Type Definition (DTD) in order to become well-formed and valid. JSON is a lightweight data interchange format for representing objects.

- **Really Simple Syndication (RSS) or Atom Feeds:** An extension of XML, allows the syndication, aggregation and notification of data. The feed can contain headlines, full text articles, summaries, metadata, data and various multimedia.

- **Simple API for XML (SAX), Document Object Model (DOM) and Extensible Stylesheet Language (XSL):** Not only is storage and distribution of data important, but so is the ability to extract useful information from the data. For this, both the data and multiple schema/DTD which define the data are required. SAX and DOM are Application Program Interfaces (APIs) for inspecting the entire contents of the data. XML Path Language (XPATH) and XML Query Language (XQUERY) act as filters designed as XSL which transform the XML document and allow specific queries.

- **Asynchronous JavaScript and XML (AJAX), Adobe Flex, JavaFX and Microsoft Silverlight:** Allowing development and deployment of cross platform rich Internet applications with immersive media and content. These applications utilize Remote Method Invocations (RMI) and Remote Procedural Calls (RPC) to servers to allow distributed interprocess communications. Web 2.0 application layer

protocols that allow this functionality include Simple Object Access Protocol (SOAP), Representational State Transfer (REST) and XML-RPC.

– **Mashups:** The merging of content from different sources, both client- and server-side.

CI is based on derived intelligence extracted from explicit and implicit user generated data, and therefore data representation is a core component for CI. XML is recommended by the World Wide Web Consortium (W3C) and is a fee-free open syntax which can be used to share information between different kinds of computers, different applications, and different organizations. This openness is highly important because it allows accessible-by-all data without needing to pass through many layers of conversion. Without XML, core components of Web 2.0 technologies would not be able to collaborate and achieve CI. The list of collaborating technologies exchanging information in XML is constantly varying - which reflects the precise character of the perpetual-beta. Current XML-based technologies which can be used for CI include Web Services Description Language (WSDL), Web Ontology Language (OWL), Linguistics Markup Language (LGML), Attention Profiling Markup Language (APML), Geography Markup Language (GML), and Predictive Model Markup Language (PMML). The languages defining various information can each be separated into the groups of: collaboration-based, explicit-based, implicit-based and intelligence-based. Section 3 describes each of these components working together for CI. The following sub sections briefly introduce each of these technologies which are then combined into the GKD from the Geoweb for CI process.

2.1 Web services description language (WSDL)

The first part of collaborating services is to provide a way for the services to communicate and describe what services they offer. The Services Description Language Version 2.0 (WSDL 2.0), is a W3C recommended XML language for describing Web services. The WSDL describes Web services in two fundamental stages. The first being abstract or document driven, which describes a Web service in terms of the messages it sends and receives; messages are described independent of a specific wire format using a type system, typically XML schema. The way messages are exchanged defines an operation which is defined by a message exchange pattern which identifies the sequence and cardinality of messages sent and/or received as well as who they are logically sent to and/or received from. The second stage defines the concrete or procedural-oriented level of the service, which defines how a service accepts bindings and associates with network endpoints, or ports [5]. The data exchanged by the Web service are defined as elements and are described with a unique name, and data type. Elements can be of simple types, complex types or be defined in an XML Schema Definition (XSD), DTD, REgular LAnguage for XML Next Generation (RelaxNG) and Resource Description Framework (RDF) file.

2.2 Web services choreography description language

The Web Services Choreography Language (WSCL) is a W3C candidate recommendation targeted for composing interoperable, peer-to-peer collaborations between any type of participant regardless of the supporting platform or programming model used by the implementation of the hosting environment [8]. The WSCL is a collection of components which builds an architecture stack targeted for integrating interacting applications which consists of:

– Defining the basic formatting of a message and the basic delivery options (SOAP);

– Describing the static interface and data types of the Web service end points (WSDL);

– Allows publishing the availability of a Web Service and its discovery from service requesters (Registry);

– Allows authentication of participants to ensure that exchanged information are legitimate and not modified or forged (Security layer);

– Allows reliable and ordered delivery between participants (Reliable Messaging layer);

– Allows the use of protocols for long-lived business transactions and enables participants to meet correctness requirements (Context, Coordination and Transaction layer);

– Describes the execution logic of Web services and rules for consistently managing non-observable data (Business Process Languages layer);

– Defines a common viewpoint of the collaborating participants describing their complementary observable behavior (Choreography layer);

The draft insists that the future of E-Business applications requires the ability to perform long-lived, peer-to-peer collaborations between the participating services, within or across the trusted domains of an organization. The WSCDL is the means by which technical multi-participant contracts can be created and viewed from a global perspective.

2.3 Attention profiling markup language (APML)

The Attention Profiling Mark-up Language (APML) is an XML-based portable file format containing a description of the user's rated interests. The APML also attempts

to contain other forms of attention data such as Attention.XML, Instant Messaging (IM) conversations, browser history, emails and other documents. The APML promises to make it easier for Web services to collect attention information of individual users to cater for the needs of individual and general users. The most compelling reason for the adoption of APML is that it defines an open and public standard of profiling that the user has direct access to. This means the user can directly be aware of what information is being shared about them and certain that Web services can provide exactly what they want. This differs from traditional captured user information by companies which tends sometimes be regarded as private and sacred. Attention information is kept up-to-date because APML is a lossy format, which maintains only the current trends and styles of the user.

2.4 The semantic web: web ontology language (OWL) and resource description framework (RDF)

The OWL and RDF are considered as the core technologies underpinning the Semantic Web; a collaborative effort led by W3C with participation from a large number of researchers and industrial partners with the aim to separate data from specific applications and making it possible for the web to understand and satisfy the requests of people and machines to use the Web content. The Semantic Web is not only concerned about the integration and combination of data drawn from diverse sources, but also how the data relates to real world objects so that both people and machines may understand and analyze the data on the Web. The OWL and RDF achieve this by publishing in languages specifically designed for data rather than just documents and the links between them. The network of linked data has been described as the Giant Global Graph (GGG), as opposed to the HTML-based World Wide Web (WWW) [4].

The OWL is designed for use by applications that need to process the content of information instead of just presenting information to humans [3]. OWL 1.0 is currently a W3C recommendation and is currently being updated to OWL 2.0 though a working draft. On top of the features of OWL 1.0, OWL 2.0 is designed to facilitate ontology development providing classes, properties, individuals, and data values stored as Semantic Web documents, with the ultimate goal of making Web content more accessible to machines. The Multimedia Web Ontology Language (MOWL) is a further refinement by the W3C which has been designed to facilitate semantic interactions with multimedia contents. The MOWL was also merged with the Knowledge Description Language (KDL) to allow semantic processing of media data calls for perceptual modeling of domain concepts with their media properties. A further extension of MOWL allows semantics for spatio-temporal relations across media objects and events. The OWL and MOWL are most commonly serialized using RDF/XML syntax.

The RDF is a W3C recommended extension and revision of XML for conceptually describing and modeling information implemented in web resources. The fundamental aim is to identifying information using Web identifiers (using Uniform Resource Identifiers, or URIs), and describe it in the form of a subject-predicate-object triple expression so that machine intelligence can store, exchange, and use machine-readable information distributed throughout the Web. The information is represented as a graph of nodes and arcs, with each node being referenced by a unique URI. This allow data to be processed outside the particular environment in which it was created, in a fashion that can work at Internet scale [9]. The triple describes the relationship of the subject and the object of the information given the conditional predicate. An example from the W3C RDF primer describes the statement: "http://www.example.org/index.html has a creator whose value is John Smith", as the following RDF statement:

- a subject http://www.example.org/index.html;
- a predicate http://purl.org/dc/elements/1.1/creator;
- and an object http://www.example.org/staffid/85740.

The URI references are used to identify not only the subject of the original statement, but also the predicate and object, instead of using the words "creator" and "John Smith", respectively [1]. Another particular format might be more direct and easily understood, however the RDF's generality and potential for collaborative intelligence through sharing gives it great value. Another advantage to the RDF is the URIs can define real locations of the referenced information. In this sense, the RDF can also provide a means for geospatial indexing of the information which can be used by the GKD process to identify particularly interesting patterns.

2.5 Geography markup language (GML)

GML serves as a modeling language for geographic systems as well as an open interchange format for geographic transactions on the Internet. It is an extension to XML that allows the ability to integrate all forms of geographic information (discrete, areal and sensor) onto data. It does this by allowing a rich set of primitives that include features, geometry, coordinate reference system, time, dynamic features, coverage, unit of measure and map presentation styling rules. The way that data is represented by GML is defined by a GML profile namespace which defines restricted subsets of GML. These profiles can be built on specific GML profiles or use the full GML schema set. The GML can be used as a standalone data format or be included as an extension to other XML-based formats to give added spatial dimensions.

2.6 The predictive model markup language (PMML)

The Predictive Model Markup Language (PMML) is an application and system independent interchange format for statistical and data mining models [18]. It is an XML-based language developed by the Data Mining Group (DMG) and allows models to be created within one vendor's application, and use other vendors' applications to visualize, analyze, evaluate or otherwise use the models. Previously, the exchange of fully trained or parameterized analytic models was very difficult, but PMML allows effective utilization between applications and is complementary to many other data mining standards. The PMML also defines the input and output format of data and how, in terms of standard data mining terminology, to interpret their results. This kind of intelligence sharing is critical between collaborating CI servers and clusters and allows for an ensemble of different models which can be used to increase the accuracy of classification [10].

3 Geographic Knowledge Discovery for Collective Intelligence

3.1 Knowledge discovery process

Knowledge discovery is the process of automatically searching large volumes of data for patterns that can be considered knowledge about the data. It is often described as deriving knowledge from the input data that can be used for further usage and discovery in the process. The process generally consists of several steps that can be executed in a non-linear order. The generic steps include:

- **Data Selection**: Creating a subset of the total data that focuses on chosen foci for concentrating the data mining activities.
- **Pre-Processing**: The cleaning of the selected data to remove noise, eliminate duplicate records, filling in missing data fields and reducing both the dimensionality and numerosity of the data in order to build and an efficient representation of the information space.
- **Data Mining**: The attempts to uncover interesting patterns.
- **Interpretation and Reporting**: The evaluation and attempted understanding of the results of the data mining process.
- **Utilization**: The use of the learned knowledge to provide accurate decision support for the utilizing industry.

Data mining is an ongoing popular research topic that focuses on the algorithms for revealing hidden patterns and information in the data. These include segmentation, dependence analysis, deviation and outlier analysis, regression and cluster analysis [7, 13]. The possible types of features of the data can be nominal, ordinal, interval-scaled, ratioed and any combination of all these types. Depending on the type of data, a distance metric is used to measure similarity and dissimilarity between the objects. By comparing these similarity and dissimilarity metrics, interesting patterns can be found within the data [6].

3.2 Collaborating web services for CI

In order to perform GKD for CI, a well formed system must be established in order to coordinate contributing services. Using Web 2.0 technologies and the Semantic Web, we define a process whereby Web services may share information in order to improve decision-making. This process is defined in Figure 2. A Web service can collaborate

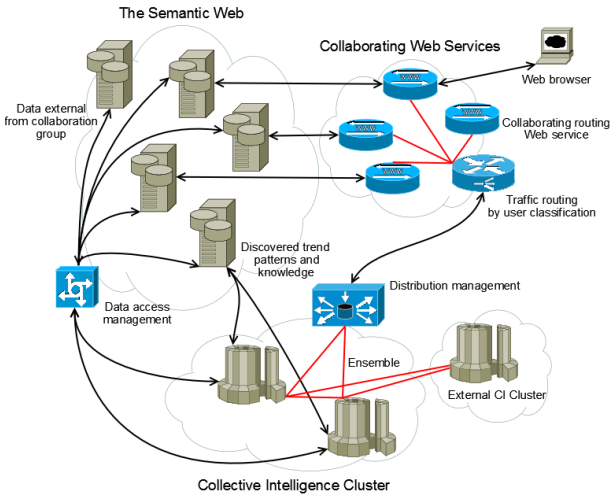


Figure 2: Collaborating Web services for CI.

with others in various ways. If the users of a transaction-based Web service are willing, then information regarding the users, products and transactions can be shared and used within the GKD process to discover patterns and knowledge for effective decision-making. However, a Web service can also not contribute information and yet still collaborate to achieve effective decision-making. These kinds of Web services interact with the discovered patterns and knowledge to route to the contributing transaction-based Web services. If traffic is routed from a collaborating Web service to an eventual profitable action, then both Web service share the profits. In this configuration, a Web service can still make a profitable action when a user does not initiate a transaction with them by effectively routing the user to a collaborating partner. A non-transaction-based Web service can profit by being popular among Web users and routing profitable traffic to transaction-based Web services. How the user is routed is determined by the discovered patterns and knowledge by the GKD process. In order for a Web service to begin collaboration, the WSCDL is used for

determining Web service information which can consist of:

- Messaging format and service end-points agreement with WSDL;
- If contributing information or purely routing-based;
- Determine security rules and how to access information provided (if any);
- Determine business rules and action of successful profitable traffic.

Information being gathered by Web services on the current user can be given to the traffic router which classifies the user and provides routing to potentially profitable destinations. The information can be anything from user gathered details, shopping basket analysis or blog, forum, tagging and rating analysis to determine APML-based information. How the user is classified is determined by the previously discovered patterns and knowledge by the CI clusters running the GKD process.

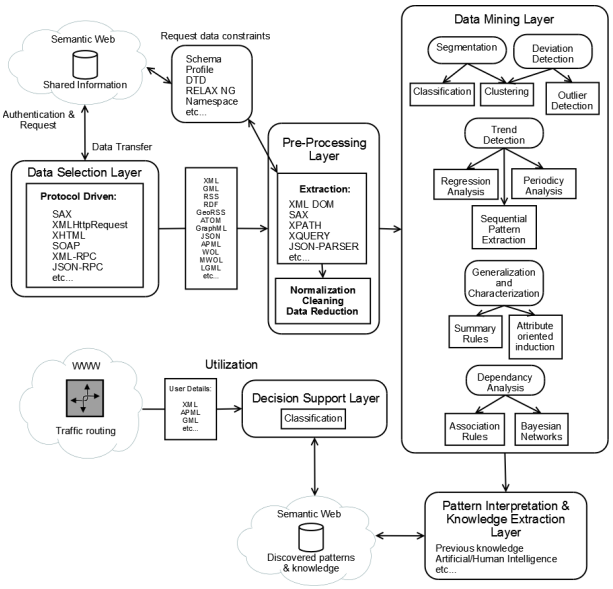


Figure 3: GKD framework for CI.

3.3 GKD for CI

Figure 3 describes GKD framework for CI. The GKD process produces useful patterns and knowledge from data retrieved from the Geoweb. The data does not just come from the collaborators, but also from publicly available linked semantic information via RDF and information collected via Web crawlers. The power behind patterns being discovered from diverse sources is that they represent global trends, as opposed to finding local trends from a singular source. The greater the diversity and number of sources - the greater the accuracy of user classification and decision support. The CI cluster is made up of multiple local CI servers with possible connections to external CI servers. This configuration is an ensemble method which aims to increase the accuracy of classification at the expense of increased complexity [10].

The first stage of the GKD process is the acquisition of data. The data is segmented and processed in various ways depending on the data mining model being used. What is common among all methods is the type of data which is available from the Semantic Web, which will be some form of XML or JSON data. When the WSCDL is used to setup collaboration, the Web service data formats and end-points are detailed so the CI server knows exactly what it is retrieving. Data constraints are determined by requesting the schema/profile information from the Web services. This allows the data to be extracted into a refined information space suitable for the data mining layer. The data mining layer can be one or a combination of models. New data mining models are constantly being discovered and refined and it is important that this layer be modular in order to easily adapt to changes.

Discovered patterns are analyzed and the models determined as novel and potentially useful are stored back onto the Semantic Web. The patterns can be stored in any format deemed suitable, but there are quite a number of advantages of producing results in a format which can be easily shared between applications, such as the PMML. This is especially important with external CI cluster collaboration when the discovered models must be compared in order to achieve greater accuracy. The discovered models represent current trends within analyzed data and are constantly changing. The advantage of determining trends from multiple sources is that certain trends will become prevalent in certain areas before other areas. Once a new trend is detected then collaborating Web services can take advantage of this information and maximize the potential profitable traffic to the trend. Web services can also submit user tracking information, such as APML-based information which can be used with the pre-determined models to classify the user to a certain group. This allows user specific traffic routing which again can greatly increase the chance of profitable traffic.

However, traffic routing need not only be Web based. Decision support can also make use of the increasing amounts of spatially reference information in order to determine real-world geospatial routing. If a user is classified as a particular group and their geospatial location is known then real-world profitable traffic can be achieved by suggesting Web services associated with the real-world elements of their group. An example of this would be to classify a particular user as a fan of a football club and to suggest products and the location of sports stores from Web services associated with that particular football club. Another possibility would be to offer current trend information related to the football club in order to deepen the relationship with the club which may eventually lead to profitable traffic. Another example scenario would be to determine that a user is interested in a particular food group and

to offer information and links to nearby restaurants of that food group. With the addition of spatial dimensions, recommendations to the user can take on new aspects and be represented as real locations on a Web map. This is the real power of using the GKD process to aid decision making. The collaborated data from the Semantic Web can then also be seen as a spatially indexed Geoweb which can be used for segmentation queries to determine potential profitable traffic for the classified user. The geospatial information can even play an important part when generating the models for classification; depending on various trend regions. The possibilities for profitable traffic from GKD from collaborating Web services is literally endless.

4 Case Studies

The following subsections give case studies using the GKD for CI process to demonstrate the potential usefulness of the system. In this study, we utilize the generalized Voronoi diagram for space tessellations and clustering for user profile segmentation. Datasets are retrieved from various mashups and visualized with mapping websites.

4.1 Restaurant recommendation case study

In this case study, let us assume a food and wine information Web service has been recording a user and building a profile using the APML. The Web server records the user searching for Château Pétrus information and sends the APML to the CI cluster for recommendations to suggest to the user. The APML contains the information related to the Château Pétrus in the RDF format and we are able to retrieve machine understandable information via its URI. With the aid of the Semantic Web, it determines that Château Pétrus is a beverage originating from France which is consumed by humans usually when dining out. The APML also contains location information relating to the user in the GML. The CI server processes these attributes with learned models to determine matches to products of the collaborating Web services that are near the user's location. It determines a number of wine distributers and restaurants which have the Pétrus in stock. Details of the distributers and restaurants which include stock number, price and location information are returned back to the Web service and used to generate the Web map as shown in Figure 4. To entice the user further, the average ratings and small snippets of reviews can be added to each location on the Web map. If the user follows any of the suggestions and either buys from a distributer or books at one of the restaurants, then the distributer or restaurant profit from the sale and the originating food and wine information Web service receives a portion from the total profit. This case shows how Geoweb and the Semantic Web are able to connect online interactions to result in real world transactions.

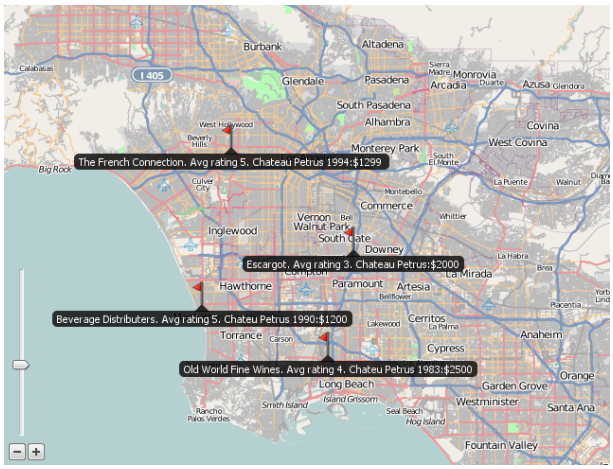


Figure 4: Recommended locations for Château Pétrus: including average rating and prices.

4.2 Recommending discovered trends

In this case study, let us assume that the CI clusters are updating their models as well as determining emerging, continuing, and fading trends from the current conditions on the collaborated Web services information. A trend is considered emerging when it has been newly detected for the current update period; continuing if it still remains from the previous period; and fading if the trend is no longer detected. The detection of trends can be useful for maximizing the possible profitable traffic to the Web services associated with current trends. The trends can be detected by the increase of sales for a particular product or even by digesting information from contributed user information to blogs, news and review Web services which might come from many various locations. Let us assume that the CI cluster have discovered an emerging trend of Brazilian coffee in the area of New York from a subset of collaborating Web services. The CI cluster finds associations of the collaborating Web service data with Brazilian coffee and updates the learned models. Now let us assume that a user is searching for classy coffee shops around New York. This tracked information is sent from a Web service to the CI cluster which determines, through association pattern mining, that classy coffee shops are linked to popular coffee. The CI server classifies popular coffee in New York as Brazilian coffee, which was pre-determined as an emerging trend. The CI server then searches the collaborated data for coffee shops with Brazilian coffee and returns a list of matches back to the original Web service which are nearest to New York, which can be generated as the Web map described in Figure 5. Using determined trends as recommendations can increase the probability of profitable traffic. The ability for GKD to determine trends for real locations from spatially referenced data and model them effectively to the user are the results of the Geoweb and Semantic Web information.

The known user location is rather vague and the exact location cannot be determined. To overcome this, the results

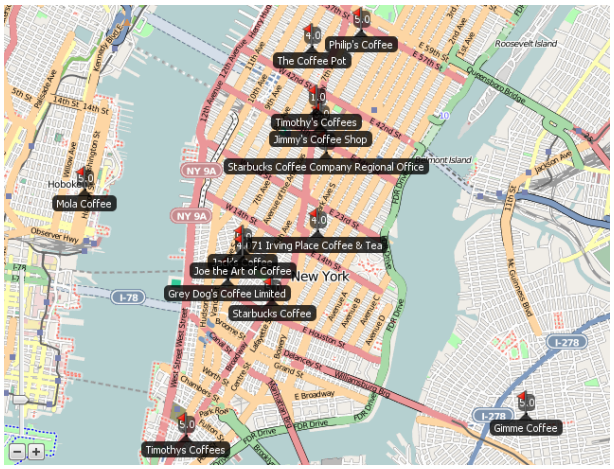
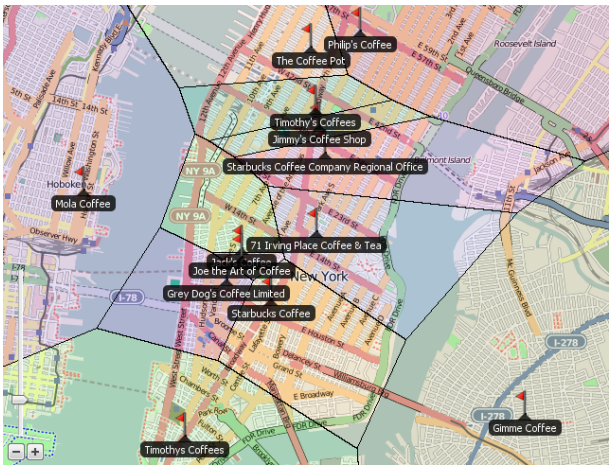
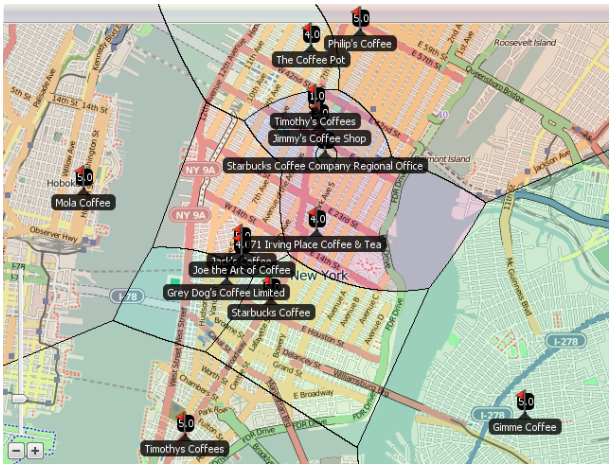


Figure 5: Recommended coffee shops with ratings.



(a)



(b)

Figure 6: Examples of recommendation links displayed as locations on a Web map: (a) Recommended coffee shops with segmentation; (b): Recommended coffee shops with weighted segmentation based on user ratings.

can undergo Web map segmentation; an effective visualization technique to help aid decision support. The results of this are shown in Figure 6 (a). Web map segmentation is not just limited to ordinary distriction but can also produce many various visualizations which could be used to further entice the user. Figure 6 (b) shows average user ratings used to create weighted regions associated with the coffee shops. The generalized Voronoi diagram has been used for space tessellations [15]. The user need only to determine which region they are located in to decide the closest highly rated coffee shop. Added spatial dimensions to information can be used to greatly enhance the depth of the information. This kind of specific information might just be what users require in order to engage them into profitable actions.

5 Final Remarks

In reality there could be many different number of the cases described in Section 4 because GKD from the Semantic Web produces extremely versatile patterns and models which can be used to determine potentially profitable traffic in a vast number of ways. In order for a true collaborative infrastructure to exist, Internet developers must try and implement their Web services using Web 2.0 technologies which conform to W3C and Open Geospatial Consortium (OGC) protocols to build the required foundation which collaborating Web services can exist on. At the moment, the Internet still consists of dominantly network as information content. There exists great potential for the increase of profitable traffic with the more collaboration that is achieved. However, there are still many components related to CI and GKD which can still be improved.

Because Web 2.0 content is forever changing and increasing, GKD techniques must be developed that can handle diverse data types which does not only consider the size of the data - but also the throughput which streaming information must be processed. A user interaction with a Web service may be near instantaneous - but the same is not necessarily so for the processing required to analyze and update current models produced by GKD. Better techniques into producing dynamically updating models under heavy streaming loads needs to be explored in the future. However, the processing time is not the only issue related to discovering models. Current GKD techniques are still relatively new and considered as an emerging research field. How can new techniques be made that can cope with the extremes of massive streaming Web data [7]?

Another problem which does not focus on the technical issues is the one regarding privacy [12]. Researchers need to ask themselves if discovering new knowledge about individuals is breaching ethical privacy. We cannot observe people going to work, seeing what they do, what they like to buy, how they invest their money, finding about their personal views without their permission. Is it okay to use this same kind of information about people that is distributed on the Internet? However, users who share private information about themselves allows target marketing with a higher accuracy - which can benefit the user because it allows them to get exactly what they want. In the future, users and Web services must be allowed the freedom to collaborate with-

out the fears of breaches of privacy and laws. This means collaboration technologies must be designed to easily allow collaboration while also circumventing fraudulent activity.

References

[1] Rdf primer. http://www.w3.org/TR/2004/REC-rdf-primer-20040210/, Februar 2004. Stand: 15.4.2009.

[2] S. Alag. *Collective Intelligence in Action*. Manning Publications, 2008.

[3] Sean Bechhofer, Frank van Harmelen, Jim Hendler, Ian Horrocks, Deborah McGuinness, Peter Patel-Schneijder, and Lynn Andrea Stein. OWL Web Ontology Language Reference. Recommendation, World Wide Web Consortium (W3C), February10 2004.

[4] Tim Berners-Lee. Giant global graph. Blog, 11 2007.

[5] Roberto Chinnici, Jean-Jacques Moreau, Arthur Ryman, and Sanjiva Weerawarana. Web services description language (wsdl) version 2.0 part 1: Core language. World Wide Web Consortium, Recommendation REC-wsdl20-20070626, June 2007.

[6] V. Estivill-Castro and I. Lee. Argument Free Clustering via Boundary Extraction for Massive Point-data Sets. *Computers, Environments and Urban Systems*, 26(4):315–334, 2002.

[7] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, San Francisco, C.A., 2000.

[8] Nickolas Kavantzas, David Burdett, Greg Ritzinger, Tony Fletcher, Yves Lafon, and Charlton Barreto. Web services choreography description language version 1.0. World Wide Web Consortium, Candidate Recommendation CR-ws-cdl-10-20051109, November 2005.

[9] Graham Klyne, Jeremy J. Carroll, and Brian McBride. Resource description framework (RDF): Concepts and abstract syntax. W3C Recommendation, Feb 2004. Available at: http://www.w3.org/TR/rdf-concepts, last access on Dez 2008.

[10] L. I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley, New Jersey, 2004.

[11] I. Lee, K. Lee, and C. Torpelund-Bruin. Voronoi Image Segmentation and Its Application to Geoinformatics. *Journal of Computers*, 4(11):1101–1108, 2009.

[12] Y. Lindell and B. Pinkas. Privacy Preserving Data Mining. *Journal of Cryptology*, 15(3):177–206, 2002.

[13] H. J. Miller and J. Han. *Geographic Data Mining and Knowledge Discovery: An Overview*. Cambridge University Press, Cambridge, UK, 2001.

[14] Harvey J. Miller. *Geographic Data Mining and Knowledge Discovery*. Handbook of Geographic Information Science, 2004.

[15] A. Okabe, B. N. Boots, K. Sugihara, and S. N. Chiu. *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*. John Wiley & Sons, West Sussex, second edition, 2000.

[16] Tim O'Reilly. What is web 2.0? design patterns and business models for the next generation of software., 2005.

[17] V. Podgorelec, L. Pavlic, and M. Hericko. Semantic Web Based Integration of Knowledge Resources for Supporting Collaboration. *Informatica*, 31(1):85–91, 2007.

[18] Stefan Raspl. Pmml version 3.0 - overview and status. In *KDD-2004 Workshop on Data Mining Standards, Services and Platforms (DM-SSP 04), KDD-2004 The Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004.

[19] I. Svetel and M. Pejanovic. The Role of the Semantic Web for Knowledge Management in the Construction Industry. *Informatica*, 34(3):331–336, 2010.

[20] C. Torpelund-Bruin and I. Lee. Geographic Knowledge Discovery from Geo-referenced Web 2.0. In *Proceedings of 2008 International Workshop on Geoscience and Remote Sensing*, pages 291–294, Shanghai, China, 2008. IEEE Computer Society.

[21] C. Torpelund-Bruin and I. Lee. When Generalized Voronoi Diagrams Meet GeoWeb for Emergency Management. In H. Chen, C. C. Yang, M. Chua, and S-H. Li, editors, *Proceedings of the Pacific Asia Workshop on Intelligence and Security Informatics*, Lecture Notes in Computer Science 5477, pages 64–75, Bangkok, Thailand, 2009. Springer.

[22] H. Wang, X. Jiang, L-T. Chia, and A-H. Tan. Wikipedia2Onto. Building Concept Ontology Automatically, Experimenting with Web Image Retrieval. *Informatica*, 34(3):297–306, 2010.