Cognitive Location-Aware Information Retrieval by Agent-Based Semantic Matching

Eddie C. L. Chan, The Hong Kong Polytechnic University, China
George Baciu, The Hong Kong Polytechnic University, China
S. C. Mak, The Hong Kong Polytechnic University, China

ABSTRACT
This paper proposes semantic TFIDF, an agent-based system for retrieving location-aware information that makes use of semantic information in the data to develop smaller training sets, thereby improving the speed of retrieval while maintaining or even improving accuracy. This proposed method first assigns intelligent agents to gathering location-aware data, which they then classify, match, and organize to find a best match for a user query. This is done using semantic graphs in the WordNet English dictionary. Experiments will compare the proposed system with three other commonly used systems and show that it is significantly faster and more accurate.

Keywords: Agent, Information Retrieval, Location-Aware, Semantic, TFIDF

INTRODUCTION
One of the most challenging problems in retrieving the location-aware information is to understand the behavior of users and how it suits the current location. Wireless tracking applications are popular ways to help user navigate that may make use of current location-aware information. However, very often, the information retrieved in common search engines is both excessive and unstructured from the user’s point of view. The basic requirement of an effective location-aware retrieval system is that it should match user queries and provides accurate information where users access location-aware (e.g., pervasive computing-enabled) applications and services. The retrieval system should do this in an organized and efficient way.

In recent years, researchers have focused on how to provide higher accuracy and faster retrieval by making use of keywords and textual semantics. However, the results so far have been unsatisfactory. First, the truth-conditional semantics that are often applied provide only a very limited account of meaning. Second, information is derived from few sources and the information from those sources is not structured. In these circumstances, the value of the

DOI: 10.4018/jssci.2010070102
information that can be extracted is very limited. Third, information collection is an expensive, time-consuming process that is often carried out manually. This makes it very difficult to build, maintain and grow comprehensive databases. Forth, some approaches, such as Naïve-Bayes (Danesh et al., 2007) and K-Nearest Neighbors (K-NN) classifiers (Weiss et al., 1996) machine learning approaches ignore the semantic meaning in text classification. This leads to inadequate search results. Finally, location-aware information is distributed in different locations and some information, traffic information for example, they will quickly become out-of-date so location-aware information must be updated frequently.

This paper proposes an agent-based semantics retrieval system for location-aware information. The proposed system uses the WordNet (WordNet, 2004) dictionary to construct the semantics graph structure of a location. This graph structure sets the weight values between edges and nodes (words) of the semantics and the classic information retrieval, term-frequency-time-inverse-document-frequency (TFIDF) technique is modified with semantics weight values. As the location-aware information is distributed in different locations, the approach also implement agents using the IBM Agent Software Development Kit (IBM, 2004). These mobile network agents are programs that can be dispatched from one computer and transported to a remote computer for execution. Arriving at the remote computer, they present their credentials and obtain access to local services and data. The remote computer may also serve as a broker by bringing together agents with similar interests and compatible goals, thus providing a meeting place at which agents can interact. The proposed system uses four types of agents: one each to gather, classify, match, and organize information.

The proposed system offers a number of benefits. First, the proposed semantics graph structure provides a hierarchical structure for the location-aware information. Second, the proposed agent system obviates the need for extensive manual information grasping. Third, the modified TFIDF with semantics weights improves the accuracy of matching keywords and provides a more meaningful result from the point of view of semantics. Forth, the agent can update information directly by communicating with its neighbor agents. Finally, it is fast and cost-effective.

The rest of the paper is organized as follows: Section 1 presents the related work of information retrieval techniques and agent-based information systems. Section 2 describes cognitive semantics TFIDF technique in our system. Section 3 presents the system implementation and architecture of agent-based cognitive semantics retrieval system. Section 4 describes the experiment setup and result. Section 5 presents the case study. Finally, Sections 6 offers our conclusion and future work.

1. RELATED WORK

In this section, we summarize current research works of the information retrieval techniques, agent-based information retrieval systems and positioning system in the area of cognitive informatics and cognitive computing (Baciu et al., 2009; Wang, 2003, 2007, 2009).

1.1 Information Retrieval Techniques

There are a number of information retrieval techniques. These include, Naïve Bayes classifier (Lewis & Ringutte, 1994), linear or quadratic discriminated analysis (LDA or QDA) (Hull et al., 1996), artificial neural network (Tan et al., 2006), back-propagation-based technique (Nawi et al., 2008) and C4.5 (Ruggieri, 2002) rule learning technique. Naïve Bayes classifier (Lewis & Ringutte, 1994) suffers from its assumption that words are independent of each other and are, therefore, less accurate than a more complex model. While standard parametric methods such as LDA, QDA (Hull et al., 1996) and regression are mostly criticized as being dependent on strong assumptions.
about the distribution of the underlying data, classification and pattern recognition methods require large number of training points. Artificial neural network-based (Tan et al., 2006) approaches are blamed to be black-box models thus not being able to provide insight into the complex interactions of the ecosystem processes, although they are able to overcome the difficulties associated with traditional statistical models. Back-propagation-based techniques (Nawi et al., 2008) are inherently off-line, that is iterative, methods using all the available data at once. This techniques need to retrain the entire database again in order to get each new observation, thus requiring a significant amount of computational resources and time. The motivation in C4.5 (Ruggieri, 2002) is to optimize rules locally by dropping conditions, as the initial rule set, being generated from a decision tree, is unduly large and redundant. But the process itself is rather complex and heuristic.

1.2 Agent-Based Information Retrieval Systems

Agents such as Harvest (Brown & Danzig, 2004), FAQ-Finder (Doorenbos et al., 1997), Information Manifold (Levy et al., 2005), OCCAM (Kwok & Weld, 1996), and Parasite (Spertus, 1997) rely either on pre-specified domain specific information about particular types of documents, or on hard-coded models of the information sources to retrieve and interpret documents. The Harvest system (Doorenbos et al., 1997) relies on semi-structured documents to improve its ability to extract information. For example, it knows how to find author and title information in Latex documents and how to strip position information from postscript files. Harvest neither discovers new documents nor learns new models of document structure. Similarly, FAQ-Finder (Doorenbos et al., 1997) extracts answers to frequently asked questions (FAQs) from FAQ files available on the web. In a conclusion, the performance of current agent-based information retrieval systems is not satisfied.

1.3 Positioning Systems

The Global Positioning System (GPS) (Taheri et al., 2004) is currently the standard for location sensing in outdoor wireless environments. Many GPS tracking applications has been developed and launched to the market to help users to locate themselves, such as a car navigation system and GPS in iPhone devices. Over the last several years, wireless local area network (WLAN) has evolved rapidly. Positioning systems for indoor areas (Prasithsangaree et al., 2002; Jan & Lee, 2003) using the existing wireless local area network infrastructure have been suggested. Although we have a full positioning coverage application across topologically varied landscape, recent applications ignores the importance of information driven by a location and it seems to be lack of research to study the location-aware information. The application should allow users to access or even proactively provide location-aware information where users trigger positioning function. A location-aware information is usually very useful for the user to know and understand a particular region, such that user could know where to access facilities.

2. COGNITIVE SEMANTICS

TFIDF TECHNIQUE

Normally, a location-aware application receives a keyword query from users. This section begins by describing how to use the keyword feature selection technique. It then introduces the TFIDF technique for forming a space vector model. The WordNet (WordNet, 2004) English dictionary is then used to form a graph of the semantic structure of a term/word and to set weight values to each edge of the graph. Instead of using Euclidean distance directly, the approach calculates the distance including the weight value of each vector. Finally, the information is clustered using K-NN algorithm.

2.1 Feature Selection Technique

Feature selection is a common information retrieval technique. Usually, there are three
feature selection criteria: 1) pruning of infrequent features, 2) pruning of high frequency features and 3) Choosing features, which have high mutual information with the target concept.

The first step is to prune the infrequent words from my training sets of data. This removes most spelling errors and speeds up the later stages of feature selection. The next step is to prune the most frequent words. This technique should eliminate non-content words like “the”, “and”, or “for”. Finally, the remaining words are ranked using the TFIDF technique according to their mutual information with reference to the target concept/category.

2.2 TFIDF Technique

TFIDF (Wen et al., 2001; Lo et al., 2008) is an information retrieval technique commonly used in query searching. This technique formulates the significance of a term/word according to its frequency in a document or a collection of documents. This work uses TFIDF to determine the weights that are assigned to individual terms. If a term $t$ occurs in document $d$,

$$w_{di} = tf_{di} \times \log (N / idf_{di})$$  \hspace{1cm} (1)

where $t_i$ is a word (or a term) in the document collection, $w_{di}$ is the weight of $t_i$, $tf_{di}$ is the term frequency (term count of each word in a document) of $t_i$, $N$ is the total number of documents in the collection and $idf_{di}$ is the number of documents in which $t_i$ appears. The TFIDF technique learns a class model by combining document (vectors) into a vector space model.

2.3 Cognitive Semantics Graph Structure

The WordNet English dictionary provides a hierarchical representation of English words organized according to the strength of their associations in a number of defined domains, which I will refer to here as a semantics graph. WordNet defines a weight value for each relationship in a graph.

$$otf_i = tf_i \times (1+D(t_i, t_j)^2)$$  \hspace{1cm} (2)

where $t_i, t_j$ are different terms; $otf_i$ is ontology-based term frequency of $t_i$, $tf_i, tf_j$ are the term frequency respectively to $t_i$ and $t_j$; $D(t_i, t_j)$ is the depth between $t_i$ and $t_j$. $D(t_i, t_j)$ can be calculated.

For example, assume the term frequencies of “tart” and “unpleasant” are 3 and 2 respectively and depth between “tart” and “unpleasant” is 3. The ontology-based term frequency of “tart” will be $otf_i = 3 \times (1+D(t_i, t_j)^2) = 3 \times (1+1/3)^2 = 5.33$. The ontology-based term frequency of “unpleasant” will be $otf_j = 2 \times (1+1/3)^3 = 4.67$. After adjustment of each term frequency, their term frequency value may increase. The two terms could thus become more significant after computing TFIDF.

2.4 Term Document Matrix

A term document matrix (Sudarsun et al., 2006; Jaber et al., 2007) is widely used in natural language processing and information retrieval as a way to reduce space requirements and speed up searching. A term document matrix is a large grid representing every document and content word in a collection. A term document matrix is generated to store my data. (2) is used to get a new (ontology-based) TFIDF value for each term and then a document vector denoted by $A = \{a_1, a_2, ..., a_n\}$ where $n$ is the total number of documents. A document vector contains three elements, a term, an ontology-based TFIDF, and a term frequency. These elements allow each document vector to be formed as a matrix in which each row stands for a term and each column stands for a document. Each cell will contain the ontology-based TFIDF weight of a
2.5 K-NN Algorithm

After developing a set of term document matrices, the K-NN algorithm is applied to a document \( w \) and a set of document clusters \( C = D = \{d_1, d_2, \ldots, d_n\} \) is a set of distance between \( w \) and \( c_i \). It is possible to estimate the similarity between a document and a set of document clusters by calculating the Euclidean distance \( |w - c_i| \). In a hierarchical clustering, two clusters \( c_i \) and \( c_j \) are randomly chosen and their similarity \( \text{sim}(c_i, c_j) \) is calculated. They are merged if the similarity value is greater than the threshold value. Otherwise, this step repeats until reaching the termination condition with \( K \) clusters.

\[
d = \frac{\sum_{i=1}^{n} d_i}{\sum_{i=1}^{n} \frac{1}{|w - c_i|}}
\]  

3. SYSTEM IMPLEMENTATION

In the previous section, we have described how our agents use the modified TFIDF to find the suitable information. This section introduces the proposed agent-based approach to implementing the system. It uses four different types of agent, which handle information grasping, classification, matching, and organizing. (1), (2) and (3) are applied into following four agents to grasp the content. Figure 2 shows the system infrastructure of the agent-based semantics retrieval system while the following subsections describe each of the four types of agents. All agents are implemented with IBMAglets (IBM, 2004). IBMAglets provides a collaborative platform to build sophisticated application with interactive services among different modules. It also provides an asynchronous runtime environment in an agent server enables parallel processing to increase system performance,
while distributed architecture in an agent system supports operation on various scales without damaging the system performance.

3.1 Information Grasping Agent

The information-grasping agent consists of two sets of agents, one static and the other dynamic. The message (information) stored in this agent is based on a XML structure. The static information-grasping agent is installed in a server to collect location-aware information from URL websites, such as restaurant and tourism websites. Static agents are always online and maintain a 24-hour record of RSS (rich site summary) information from popular websites.

Dynamic information grasping agents collect information in a mobile environment. In most of cases, they collect location-aware information through information exchanges between users or sometimes collect data invoked by other real-time systems such as surveillance systems or supply chain RFID tracking systems. Increasing the number of dynamic information agents will help keep information more up-to-date but on the other hand puts an added burden on network traffic. A user-driven process will allow only the dynamic information-grasping agent to be triggered. When the agent is invoked at a user’s request, it gathers different sources of information automatically and then updates the information with neighbors or other systems. It thus contributes to and receives the latest information through the system.

Dynamic information grasping agents has following properties: (1) Autonomous: it collects the information according to user’s preference and is not required constant human guidance. (2) Proactive: it has the ability to sense the new location and respond to collect new information stimulation appropriately. (3) Mobile: it is able to migrate itself from one host to another and bring data across different mobile platform. (4) Collaborative: it is able to make conversation with following (other) agents to achieve sociability.

Figure 2. System Architecture of Agent-based cognitive semantics retrieval system in our system
3.2 Information Classification Agent

The information classification agent classifies output information into suitable categories and calculates the Euclidean distance between the words to form a word graph, which expresses relationships between words (Figure 1). It is a core part in the system because word classification directly affects the accuracy of searching.

3.3 Information Matching Agent

The information-matching agent matches user queries and categorizes information. According to the users’ preferences, the agent analyzes and searches for information in the categorized information. The best match for the user’s requirements is found using (2) and (3).

3.4 Information Organizing Agent

The information-organizing agent organizes the search result. It provides a user-friendly interface, receiving user preferences and change the display and organization of the content.

4. EXPERIMENT SETUP AND RESULTS

This section makes use of the Chinese Learner English Corpus (CLEC) (PolyU, 2004) and in particular 500 location-related articles subdivided into six categories. There are 83 articles each for the domains Hotel, Restaurant, Traffic and Shopping mall and 84 for the domain of College and Clinic. Figure 3 illustrates the number of articles in each category.

Two sets of experiment are conducted to compare three types of techniques, the three-layer neural-network (NN), TFIDF with hierarchical clustering, and the proposed semantics TFIDF techniques. The first experiment tests the precision of categorization/clustering. The second experiment measures the computation time required to cluster using three techniques. An additional three test sets of 100 user queries is input to find out suitable information from the system. Each user query consists of at most ten words. All the experiments are conducted on a common desktop computer with a CPU.
of Intel Pentium 4.3 GHz, 2GB DDR2 SDRAM and physical storage of 200GB with 7200 rpm.

In subsection 4.1, we discuss the precision rate in three different techniques. Subsection 4.2 measures the computational time during information retrieval, and discuss the effectiveness of using different techniques.

### 4.1 Precision Rate

Table 1 and Figure 4 show the result of the precision when using the three different information retrieval techniques. The average precision across different categories of semantics TFIDF is 95.68% which is the highest of the three techniques.

#### 4.2 Speed

Figure 5 shows the computational time required to find suitable information when using three different information retrieval techniques. The NN technique takes 5295 seconds for clustering, which is approximately five times longer than the other techniques.

---

**Table 1. Comparison of the precision rate (%)**

<table>
<thead>
<tr>
<th>Category</th>
<th>Method</th>
<th>TFIDF with hierarchical clustering</th>
<th>Cognitive semantics TFIDF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NN (3-Layer)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hotel</td>
<td>68.40%</td>
<td>90.30%</td>
<td>95.13%</td>
</tr>
<tr>
<td>Restaurant</td>
<td>58.45%</td>
<td>93.37%</td>
<td>96.72%</td>
</tr>
<tr>
<td>Traffic</td>
<td>73.43%</td>
<td>93.88%</td>
<td>95.14%</td>
</tr>
<tr>
<td>Shopping mall</td>
<td>67.00%</td>
<td>94.33%</td>
<td>97.44%</td>
</tr>
<tr>
<td>College</td>
<td>70.22%</td>
<td>93.78%</td>
<td>94.38%</td>
</tr>
<tr>
<td>Clinic</td>
<td>50.23%</td>
<td>91.67%</td>
<td>95.24%</td>
</tr>
<tr>
<td>Average</td>
<td>64.62%</td>
<td>93.72%</td>
<td>95.68%</td>
</tr>
</tbody>
</table>

---

**Figure 4. Proposed Location-aware Application using Apple’s iPhone**
than the other two methods. The shortest time required finding suitable information is TFIDF with hierarchical clustering technique, at only 1138 seconds to finish clustering. The semantics TFIDF technique requires 1185 seconds.

5. CASE STUDY

In this section, we discuss our proposed system with our tourist guide application in iPhone 3GS. The system will identify the user current location with positioning technologies such as Differential Global Positioning System (DGPS) and return the favorable dinning place search result according to the user’s preference. Figure 6 shows the user interface of our application in iPhone 3GS.

There are three main layers in our iPhone application. The first layer is the positioning layer. DGPS would be used to estimate the user’s location. iPhone 3GS has already included the features of DGPS.

The second layer is the user input layer which reads users’ query, displays the restaurant information and identify the location in the Google map. We implement our system using iPhone as front-end user interface. iPhone is a touch device that it can easily control the zooming function of the map.

The third layer is our proposed agent system layer. This layer includes a location-aware database server which stores the name, address, price, type and global co-ordinate of 17,000 restaurants. We make use of the proposed system to help the user to find suitable dinning places (Table 2).

6. CONCLUSION AND FUTURE WORK

In this paper, our agent-based cognitive semantics retrieval system for location-aware information is introduced. We illustrate our cognitive semantics TFIDF technique and implement intelligent agents for information grasping, classification, matching and organizing. We also evaluate the performance of our system. Our experimental result indicates that the effectiveness of our system leads to have more accurate and faster search result. Future work will consist in user behavior model in different location and the ontology tree structure for location-aware information.

REFERENCES


Copyright © 2010, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
Eddie Chan received his BSc Degree in Computing and MSc Degree in E-commerce in the Department of Computing from The Hong Kong Polytechnic University in 2005 and 2007, respectively. Currently, he is a PhD student in the same university. His research interests include wireless communication, localization, fuzzy logic, 3D visualization of tracking system, agent technology and data mining. He is a member of the Graphics and Multimedia Applications (GAMA) Laboratory at The Hong Kong Polytechnic University.

George Baciu holds a PhD degree in Engineering and a B.Math degree in Computer Science and Applied Mathematics from the University of Waterloo. He has been a member of the Computer Graphics Laboratory and the Pattern Analysis and Machine Intelligence Laboratory at the University of Waterloo and subsequently Director of the Graphics And Music Experimentation Laboratory at The Hong Kong University of Science and Technology in Hong Kong. Currently, Professor George Baciu is Director of the Graphics And Multimedia Applications (GAMA) Laboratory in the Department of Computing at The Hong Kong Polytechnic University. His research interests are primarily in mobile augmented reality systems, user interfaces, physically-based illumination, rendering, image processing, motion tracking and synthesis for both outdoor and indoor location aware systems.

S.C. Mak received his BSc Degree in Computing in the Department of Computing from The Hong Kong Polytechnic University in 2008. His research interests include wireless communication, localization, agent technology and data mining. Currently, He is a member of the Graphics and Multimedia Applications (GAMA) Laboratory at The Hong Kong Polytechnic University.