ABSTRACT
In providing a service to mobile users, it would be critical to know what types of information they would look for in association with geo-referenced entities that may be extractable from queries or contexts. While understanding high-level user intentions in accessing the Web, such as informational, navigational, and transactional, is useful, a finer-level classification of user interests would further help adapting mobile search results to user intentions. Our research focuses on understanding what aspects of geo-referenced entities are mentioned in user queries in an attempt to create a model for user intents in geo-referenced Web searching. By collecting and analyzing geo-referenced questions posed to operational question answering systems, we delineated major aspects of non-topical information that people would seek in association with geographic information. The identified aspects were further conceptualized to develop a user interest model with three dimensions, which was validated with two sets of data. The model can be a basis for identifying user’s intent in a mobile search context as well as classifying geo-related text to be retrieved for its aspectual category.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—Query formulation, Search process; H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing—Indexing methods

General Terms
Experimentation, Human Factors

Keywords
Question Analysis, Aspects of Geo-referenced Entities, Query and Document Classification for Geo-related Text

1. INTRODUCTION
As the usage and demand of mobile search are increasing, research on understanding geo-referenced information is getting more and more attention. For searching, most recent studies are focused on matching between location names in text and those in the query. For example, ‘New York’ in a text or query needs to be disambiguated among ‘New York city’, ‘New York state’, a movie name, etc. A more difficult type of queries would handle a spatial relationship as in “a restaurant near the Eiffel Tower”.

In a mobile searching environment, however, it would be more important to know which aspect of the location the user is interested in and whether it can be identified in the potentially relevant Web pages. When a user is interested in a landmark (e.g. Statue of Liberty in New York), the interest may lie in a particular aspect such as its history, admission fees, a driving direction, a subway time schedule to the place, etc.

As a first step toward addressing the issue, we attempted to identify various aspects that are associated with geo-referenced entities in user information needs. Once turned into a user interest model, they can serve as a basis for analyzing user queries both in mobile and non-mobile search contexts and Web pages for searching. Instead of returning Web pages containing a particular geo-referenced entity, one can select those mentioning the desired aspects associated with the entity, identified from the query.

There have been several attempts to understand web search goals. Rose and Levinson [1] suggested a framework that can represent user goals based on Broder’s category [2]: Navigational, Informational, and Resource for general web search. It has two limitations for geo-referenced information search although it can be helpful for general web search. The first is that the suggested framework is too abstract to distinguish queries like “restaurant in New York” and “jogging safely in New York.” While the general goal of both queries is classified as “Informational-Advice”, they are different in their focus in the geographic location. The former is likely to look for an entity names in the location, whereas the latter focuses on a particular activity and related information, i.e., safety. In order to find different types of information or aspects of the location, it would be necessary to make a finer level distinction among the aspects.

Another limitation is that a query may not be classified to a single goal as assumed in the framework. A geo-referenced query is likely to have multiple goals as in “casual restaurant in Manhattan open on Sunday” Therefore, it is necessary to tag a page with one or more aspects tied to geo-references rather than classify them into a set of rigid categories.

To develop a finer-grained user interest model encompassing various informational aspects of geo-referenced entities, we collected and analyzed questions from a QA system, which express user’s information needs more explicitly and precisely than shorter queries generated for web search. While our ultimate goal is not to handle long questions in a traditional QA system, we chose these questions because they contain rich information for various contexts in which geo-referenced entities are mentioned. Moreover, this way of analyzing QA questions may discover more variety of user’s information needs than conducting a user study under a controlled mobile environment with the numbers of participants and situations are limited. Once we have a user interest model with various aspects, it would help processing much shorter mobile or geo-referenced search queries. The resulting model was validated with additional data obtained from another QA system.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Copyright 2010 ACM 978-1-4503-0412-2/10/11…$5.00.
2. RELATED WORK

There are a few studies that have attempted to identify user intentions in queries through a user study or a query log analysis. Expanding Broder’s categories of web queries, navigational, informational, and transactional [2], Rose and Levinson [1] proposed a new framework to represent users’ web search goals in more detailed categories. The 11 goals they delineated for web search include: Navigational, Informational-closed, Informational-open, Resource-entertainment, etc. While the framework can be well adopted for general search applications, its level is too high to be useful for identifying aspects of geo-references. In addition, a query is assumed to have only one goal for the user in this model.

Past studies for query analyses in geo-referenced information retrieval systems [4, 5, 6, 7, 8] have mostly focused on finding locations specified in a query and web documents, or discovering a spatial relationship between the location in the query such as “a restaurant near the Eiffel Tower” and a matching object. The relationship may mean that the distance between a restaurant and Eiffel Tower should be lower than a threshold. They rarely considered the user’s interest in particular aspects related to the geographical entity.

Henrich and Lüdecke [3] analyzed some characteristics of geographic information needs based on AOL queries. They identified four different aspects of geo-referenced queries. First, they defined a total of 19 object categories such as habitation, accommodation, spare time, information, etc. that are frequently used in geographic queries. For the second aspect, they categorized four predicate types: go there, gather information, buy or rent, entertain. Thirdly, they showed distributions of spatial intention such as ‘to there’ and ‘about there’. Finally, they considered a number of queries that are related to cover an area or to select a location. We posit that the web queries usually are too short to express and hence analyze the real intention of the query.

3. A User Interest Model for Geo-referenced Searching

3.1 Aspectual Categories

To develop a data-driven user interest model for geo-referenced searching, we collected 900 most popular questions from the Naver “Knowledge-In” service, which is the largest community Q&A service in Korea. Both questions and answers in various domains are provided by the casual users of the portal service. We chose the questions of the QA service rather than web queries for our study because the questions express information needs more clearly and unambiguously.

Though the dataset was collected from a local QA section, there were many questions without any reference to geographical entities because sometimes authors of questions mis-categorize them. After filtering out those impertinent questions manually, we finally collected a set of 590 questions. Three graduate students in computer science participated in analyzing the questions. They divided the set into three subsets containing roughly equal numbers of questions and annotated each set for geographic entities and the aspects of information being sought. For the first run, they had a freedom to choose their own semantic tags to describe the information need aspects. The annotated results were exchanged twice to ensure that each question be analyzed three times by the three participants.

![Figure 1. Frequency distribution of aspectual categories extracted from geo-referenced questions](image)

After agreeing to a set of annotation tags to be used, which resulted from several iterations of merging and splitting the individual tags created by the participants, they finalized the annotations of the batch given to them at the beginning. The resulting 16 annotations (or aspects of question) and their frequencies are shown in Figure 1. The meanings of the 16 aspectual categories are explained in Table 1 with examples. It should be noted that the aspects are not mutually exclusive; a question or document may contain more than one aspect.

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Meaning</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direction</td>
<td>Information about the how to get to a location</td>
<td>Direction from Insa-dong to Seoul Movie Theater on foot.</td>
</tr>
<tr>
<td>Price</td>
<td>Price for obtaining something in a place</td>
<td>What’s the ticket price for Everland (an amusement park in Korea)?</td>
</tr>
<tr>
<td>Quality</td>
<td>Quality of service or facility</td>
<td>Which hospital do you recommend?</td>
</tr>
<tr>
<td>Landmark</td>
<td>Information about a landmark in a place</td>
<td>Information about a statue of admiral Yi Sun-shin in the city?</td>
</tr>
<tr>
<td>Time</td>
<td>Admission time or opening hours</td>
<td>How long does it take to get to it?</td>
</tr>
<tr>
<td>Location</td>
<td>Where a building or place is located</td>
<td>Where is the store X located?</td>
</tr>
<tr>
<td>Reason</td>
<td>Why a place or an entity in a place has certain characteristics</td>
<td>Why is Seoul famous?</td>
</tr>
<tr>
<td>Product</td>
<td>Information about a product produced on a</td>
<td>What are the special products of Suwon (a</td>
</tr>
</tbody>
</table>
3.2 The Constructed Model

The second phase was to construct a user interest model for geo-referenced information needs by conceptualizing the 16 aspects derived from the intellectual analysis. We attempted to focus on the search intents conveyed in the questions and summarize them into a few notable dimensions or facets, so that not only questions but also answers or documents can be categorized under the scheme. Our ultimate goal is to enable search engines handle aspects of documents in conjunction with geographic entities. For example, a system should be able to identify documents containing experiences in a folk village, i.e., “experience” aspect, not general information about it often found in a typical homepage.

The resulting user interest model abstracted out of the 16 aspects is comprised of three dimensions. Each dimension has discrete values with which questions and documents can be categorized further. This three-dimensional view is not the only model on can come up with but is provided to help conceptualizing geo-referenced user interests.

3.2.1 Subjectivity

Subjectivity characterizes the degree to which subjective information is to be contained in the search result. Related aspects are mood, opinion, quality, experience, and reason. It determines whether the information is based on a fact, an interpreted fact, or a strictly experiential opinion or feeling. Fact means the user wants to know fixed or factual information such as “the bus number from Ulsan High School to Ulsan Mugeo-dong.” An example for Opinion question is “a restaurant serving delicious Kimchi in Insadong.” Interpreted fact (IF) refers to opinions that have been treated as a fact. With “what are famous attractions in Mt. Halla?”, for example, the user is asking for the general belief on something although it is not a fact. An opinion accepted or agreed by the majority of the public would belong to this category.

3.2.2 Search Target

Search Target is what the user wants to get the information about. Related aspects to this dimension are direction, products, and event. Four discrete categorical values along this dimension are: object in there, activity at there, about there, and transportation between locations. Object is a thing that exists in a location physically, such as buildings, instruments, etc. Activity indicates actions taking place or events in a place, such as jogging, dancing, drinking, etc. ‘About there’ refers to the information about a location or place excluding objects and activities. While the three categories are specific to a location, the last category, transportation, refers to inter-location information, such as “how to get to Ulsan High school from Daejeon”.

3.2.3 Restriction

Restriction indicates whether there is anything that prevents the user from physically accessing the place of interest. Related aspects are accessibility, time, number of participants, and price. These aspects may restrict user’s decision related to the place. The values for this dimension are binary: restriction and no restriction.

4. Validation

We tried to validate whether the proposed model is reasonable and useful in categorizing geo-referenced information needs via coverage and user agreement. With coverage, we attempted to see if each category is used frequently enough when it is used for aspects of can. Even though a category is found to be useful, we need to make sure that it is clearly defined enough so that a person or system can assign it to a question reliably. This can be shown by measuring the degree to which human annotators agree in assigning a category to questions.

We ran a preliminary experiment by hiring three subjects, graduate students in Computer Science. They were asked to manually annotate a random sample of 228 questions from the same dataset. From the annotated data, percentage values of the questions for which two participants assigned the same category is shown in Table 2. While two categories, Subjectivity/Fact and Search Target/Object-on-there are dominant, others appear to have sufficient coverage.

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Value</th>
<th>Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subjectivity</td>
<td>Fact</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td>IF</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td>Opinion</td>
<td>0.22</td>
</tr>
<tr>
<td>Search Target</td>
<td>About there</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>Object on there</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>Activity at there</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>Transportation</td>
<td>0.15</td>
</tr>
<tr>
<td>Restriction</td>
<td></td>
<td>0.14</td>
</tr>
</tbody>
</table>
In order to see whether the scheme works for other data, we collected a dataset from Yahoo! Answers. An added advantage is that it helps us see whether the model would be suitable for a different culture. Since Yahoo! Answers does not have a distinguishable category for geographic questions, we gathered search results with the query “NYC,” which makes most of the questions geo-referenced. A total of 300 questions were collected and annotated by three participants. The distributions of the categories show relatively even compared to the Naver KIN data as shown in figure 2. More questions in English seek for information related to Restriction, Activity at there, and Opinion. Less questions seek for Object, About there and Fact.

Having developed the model, we plan to continue our work along several directions. The first direction is to develop a text classifier for the user interest model. It is a challenging task to automatically classify pages for non-topical categories. Second, we need to apply the model to the mobile search queries rather than the Q&A questions. Since mobile search queries are too short to contain the aspects explicitly, we plan to conduct a user study for understanding possible goals and analyzing the context of search so that we see how a spatio-temporal context is related to the aspects. It is entirely possible that the model needs to be changed in the process.

6. ACKNOWLEDGMENT
This research was supported by a Microsoft Research Asia (MSRA) Faculty-Specific Project, and by the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency) [NIPA-2010-C1090-1011-0008].

7. REFERENCES