Extracting Adjective Facets from Community Q&A Corpus

Takehiro Yamamoto, Satoshi Nakamura, Katsumi Tanaka
Department of Social Informatics, Graduate School of Informatics
Kyoto University
Yoshida-Honmachi, Sakyo, Kyoto 606-8501, Japan
{tyamamot, nakamura, tanaka}@dl.kuis.kyoto-u.ac.jp

ABSTRACT

In this paper, we propose a method for helping users explore information via Web searches by using a question and answer (Q&A) corpus archived in a community Q&A site. When users do not have clear information needs and have little knowledge about the task domain, it is difficult for them to create queries that adequately reflect their information needs. We focused on terms like "famous temples," "historical townscapes," and "delicious sweets," which we call *adjective facets*, and developed a method of extracting these facets from question and answer archives at a community Q&A site. We evaluated the effectiveness of our adjective facets by comparing them with several baselines.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval

General Terms

Design, Human Factors

Keywords

Information retrieval, community Q&A, exploratory search.

1. INTRODUCTION

Recent advancements in Web search engines have enabled users to quickly obtain the information they require by issuing a single appropriate query. For example, when users want to know tomorrow's weather in Kyoto, they issue the query "Kyoto weather" to a search engine and visit the highest-ranked search results to obtain the information they require.

With some search tasks, however, users need to generate multiple queries from different aspects and perform searches iteratively. For example, if users who have never been to Kyoto want to make travel plans to visit there, they need to gather information about various aspects of Kyoto. Users searching for sightseeing spots in Kyoto might be interested in information about famous temples, places where they can enjoy leaves changing color, minor shrines, or historical castles. They also might be interested in information about famous local foods in Kyoto, restaurants that serve local foods, or authentic Japanese-sweets. Since content on the Web is constantly increasing and the information needs of people who are searching for it have become increasingly complex, the importance of exploratory Web searches has become paramount [5].

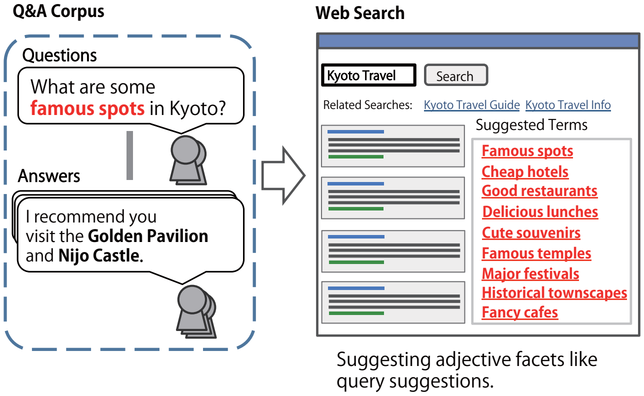

Figure 1: The concept behind our approach. We use a Q&A corpus to enhance Web searches.

However, conventional Web search engines are not sufficiently supporting users doing exploratory Web searches. The basic problem is that it is not easy for users to create queries that adequately reflect their information needs. Query suggestion [2] is a widely used technique in commercial Web search engines to assist users with generating queries. However, since query suggestion is based on the query logs of people searching the Web, it often provides excessively popular queries like "Kyoto travel guide" or "Kyoto travel information," which are not specific enough to elicit users' interests or to encourage them to browse for more information. We need to provide more appealing terms that can attract users to interests they might not have been aware of and encourage them to browse for more search results.

When users ask their friends for some information, there are various interactions between them, which is in stark contrast to using an electronic search system. In addition, there are no constraints that concern search systems. For example, when users ask one of their friends for information about traveling to Kyoto, they might ask, "I want to visit some famous temples in Kyoto. Which ones should I visit."

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.
Copyright 2011 ACM 978-1-4503-0717-8/11/10 ...$10.00.

Then, their friends might answer, “I recommend visiting the Golden Pavilion and Nijo Castle.” If we can adapt such human-human interactions to Web searches, those searching the Web can interactively explore information according to their interests and find more information than they can with conventional search systems.

In this work, we focus on community question and answer (Q&A) sites, where people post questions and answers to support users’ exploratory Web searches. The Q&A corpora that are archived in Q&A sites are caches of human-human dialogues. The key idea of our research is to use Q&A corpora to enhance Web searches, especially exploratory Web searches (Figure 1). If we can extract questioners’ interests (like “famous temples”), such knowledge can easily be applied to Web searches.

We focused on terms like *famous temples* or *delicious Japanese-sweets*, which we call the *adjective facets* of a given query, and developed a method of extracting adjective facets from questions in a Q&A corpus. The method first takes into consideration relationships between questions that contain the adjective facets and answers that contain *entities* (like “Golden Pavilion”) and then constructs a facet-entity bipartite graph, which represents the question-answer co-occurrence of adjective facets and entities. The method then applies the HITS algorithm to the bipartite graph to rank possible adjective facets. We evaluated the effectiveness of our adjective facets by comparing them with several baselines.

2. EXTRACTING ADJECTIVE FACETS

The popularity of community Q&A sites such as Yahoo! Answers and Baidu Zhidao, where people directly use natural language to post questions and answers, has recently been rapidly increasing. We will first describe and define the *adjective facets* used in our work. We will then present the core algorithms used in our method.

2.1 Adjective Facets

From our preliminary experiment, we found that people on a Q&A site are more likely to use “adjectives” than those on the Web. We focused on adjectives and extracted user’s adjective-based interests from a Q&A corpus and used them to support exploratory Web searches. However, suggesting adjectives on their own, such as “famous” or “cheap,” to Web searchers are not so useful since their meaning are quite ambiguous. In contrast, adjective-noun combinations like “famous temples” or “cheap hotels” are both more meaningful and more interesting for users. Thus, we define an *adjective facet* as a nominal phrase that matches the lexical syntactic pattern of <*adjective*> <*noun*>. The objective of this work is to extract good adjective facets that attract users’ interests from a Q&A corpus in order to supporting exploratory Web searches, like Figure 1. Our adjective facets have two characteristics in terms of supporting Web searches.

One is *unexpectedness* of terms. One important aspect of suggesting terms is to help users think up new approaches to searches. If all the terms that are suggested to users are easily generated for them, this does not effectively help them to explore information from various aspects. It is common for people to issue only nouns as queries when using Web search engines. Therefore, adjective facets, simply because they contain adjectives, are difficult for users to generate as queries and, thus, unexpected to users.

Second is *generality* of terms. For example, it is not so useful for the system to suggest term “Futaba” to users for the query “Kyoto travel” if they have little knowledge about Kyoto, since they cannot understand the meaning of the suggested term and cannot make judgments on its relevance to the query. However, it is easy for all users to understand the meaning of terms like “famous Japanese-sweets shop”. It is important to suggest terms to users that are not specific within in a certain domain in this way, especially when they do not have enough knowledge about the target domain.

2.2 Approach to Ranking Adjective Facets

We consider the question-answer co-occurrence between adjective facets and *entities* to rank adjective facets. For example, if a question is related to “famous temples” in Kyoto, the possible answers to this would intuitively contain the specific names of temples, such as “the Golden Pavilion” or “Kiyomizu temple,” that are related to the adjective facet. We called specific terms, such as the names of places, restaurants, events, people, books and so on, *entities*. Since entities can be important information that meets the questioners’ interests, the adjective facets that lead to the entities might attract Web searchers’ interests.

In this work, we made the following two assumptions:

- If important entities appear in an answer, adjective facets that appear in the question of the answer are important.
- If important adjective facets appear in a question, entities that appear in the related answers are important.

These assumptions are quite similar to those in the HITS algorithm [4], which calculates the importance of Web pages using a bipartite graph. Therefore, we consider the relationship between adjectives facets and entities as a bipartite graph and apply the HITS algorithm to the graph to rank important adjective facets.

2.3 Overview

Our method works in accordance with the following flows.

1. The method accepts query $\mathbf{q}$ made by a user.
2. The method retrieves question-answer pairs that contain query $\mathbf{q}$ in their question part. C denotes the set of retrieved question-answer pairs,
$C = \{(q_1, a_1), \ldots, (q_n, a_m)\}$, Q denotes the set of all questions in C, $Q = \{q_1, \ldots, q_n\}$, and A denotes the set of all answers in C, $A = \{a_1, \ldots, a_m\}$, and $n \leq m$.
3. The method then extracts all adjective facets $F = \{f_1, \ldots, f_{|F|}\}$ that appear in Q using a morphological analyzer. The method also extracts all entities $E = \{e_1, \ldots, e_{|E|}\}$ that appear in A. To extract entities, we used a Japanese named entity recognition module[1] and regarded the named entities that were tagged as <ORGANIZATION>, <LOCATION> and <ARTIFACT> as entities in this work.
4. After extracting adjective facets F and entities E, the method constructs a facet-entity bipartite graph G.
5. The method ranks possible adjective facets using graph G and outputs the top k ranked adjective facets. Our system then displays the obtained adjective facets to the user.

[1] http://chasen.org/˜ taku/software/cabocha/

2.4 Facet-Entity Bipartite Graph

We first construct a facet-entity bipartite graph $G = (F \cup E, \mathcal{E})$, where F and E denote the node set and $\mathcal{E}$ represents the edge set between facets F and entities E. If adjective facet $f_i \in F$ and entity $e_j \in E$ co-occur in the same question-answer pair, there is an edge between f_i and e_j.

After constructing the graph, we calculate a weight for each edge $(f_i, e_i) \in \mathcal{E}$. Let $c(f_i, e_j)$ be a weight of edge (f_i, e_j). We define $c(f_i, e_j)$ as the number of question-answer pairs in C that contain f_i in their question part and e_j in their answer part.

Let w_{ij}^{fe} be the transition probability from adjective facet f_i to entity e_j. w_{ij}^{fe} is defined as $w_{ij}^{fe} = \frac{c(f_i,e_j)}{\sum_{e_l \in E} c(f_i,e_l)}$. Similarly, the transition probability w_{ji}^{ef} from entity e_j to adjective facet f_i is defined as $w_{ji}^{ef} = \frac{c(f_i,e_j)}{\sum_{f_k \in F} c(f_k,e_j)}$.

2.5 Ranking Adjective Facets using HITS

To apply the HITS algorithm to the facet-entity bipartite graph G, we used the Co-HITS [3] framework proposed by Deng et al. Co-HITS is a general algorithm to stochastically calculate the importance of the nodes in a bipartite graph, and it contains HITS as a special case.

Let the score of adjective facet f_i be x_i and the score of entity e_j be y_j. x_i and y_j can be calculated by iterating the following two formulae:

$$x_i = \sum_{e_j \in E} w_{ji}^{ef} y_j, \quad y_j = \sum_{f_i \in F} w_{ij}^{fe} x_i \tag{1}$$

We can obtain important adjective facets on the basis of score x_i by solving Equation (1).

2.6 Query Association

Calculating Equation (1) enables us to obtain important adjective facets. However, if we simply apply Equation (1) to the bipartite graph, some irrelevant adjective facets (like “good advice”) also obtain high scores. These irrelevant adjective facets frequently appear in questions but are not related to query $\mathbf{q}$. Therefore, we incorporate the bipartite graph G with the association between query $\mathbf{q}$ and adjective facets F to remove irrelevant adjective facets.

By using the Co-HITS algorithm, we can consider the initial importance of each node in the bipartite graph. If x_i^0 is an initial importance of adjective facet f_i, Equation (1) can be modified to

$$\begin{aligned} x_i &= (1-\lambda_f)x_i^0 + \lambda_f \sum_{e_j \in E} w_{ji}^{ef} y_j \\ y_j &= \sum_{f_k \in F} w_{ij}^{fe} x_i, \end{aligned} \tag{2}$$

where $\lambda_f \in [0, 1]$ are the parameter that balances the initial scores x_i^0. If λ_f is set to 1, this equation is equal to Equation (1).

To calculate the association between adjective facet f_i and query $\mathbf{q}$, we calculate the term co-occurrence within a Q&A corpus. We used *expected pointwise mutual information* [1]. For given query $\mathbf{q}$ and adjective facet f_i, expected pointwise mutual information $\mathrm{epmi}(\mathbf{q}, f_i)$ is defined as

$$\mathrm{epmi}(\mathbf{q}, f_i) = P(\mathbf{q}, f_i) \cdot \log \frac{P(\mathbf{q}, f_i)}{P(\mathbf{q})P(f_i)} \tag{3}$$

Table 1: Categories and example queries used in the experiment.

Category	Example Query
Sports	World Cup
Science & Academics	study Mathematics
Politics & News	earthquake countermeasures
Travel	Kyoto travel
PC & Electronics	smartphone
Health	Meniere’s disease
Business & Finance	investment trust
Relationships & Life	wedding party
Education & Parenting	children discipline
Home & Food	potato recipes

To estimate probabilities $P(\mathbf{q})$, $P(f_i)$, and $P(\mathbf{q}, f_i)$, we use simple normalized frequencies: $P(\mathbf{q}) = \frac{n_\mathbf{q}}{N}$, $P(f_i) = \frac{n_{f_i}}{N}$ and $P(\mathbf{q}, f_i) = \frac{n_{\mathbf{q} \wedge f_i}}{N}$. Here, $n_\mathbf{q}$ and n_{f_i} denote the number of questions in the Q&A corpus that contain $\mathbf{q}$ and f_i, respectively. $n_{\mathbf{q} \wedge f_i}$ denotes the number of questions that contain both $\mathbf{q}$ and f_i. N denotes the total number of questions in the Q&A corpus. We set $N = 30,000,000$, which nearly equals the number of questions posted on Yahoo! Japan Chiebukuro. If the two terms frequently co-occur in the same question, the score takes a high value.

By using $\mathrm{epmi}(\mathbf{q}, f_i)$, the initial importance of adjective facet f_i can be computed as $x_i^0 = \frac{\mathrm{epmi}(\mathbf{q}, f_i)}{\sum_{f_k \in F} \mathrm{epmi}(\mathbf{q}, f_k)}$. Given the initial relevance x_i^0, x_i and y_j can be calculated by iterating the Equation (2).

3. EXPERIMENTS

3.1 Experimental Settings

To determine the effectiveness of our proposed approach, we prepared four methods.

- Frequent adjective facets in Web search results (**WEB**): This method first obtains 1,000 Web search results for given query $\mathbf{q}$ using the Yahoo! Japan Web search API. It then extracts adjective facets appear in the search results and outputs the 15 most frequent ones.
- Query suggestions (**QS**):
  This method first obtains query suggestions of given query $\mathbf{q}$ by using the Yahoo! Japan Related words API. The method then outputs the top 15 query suggestions while removing the terms contained in $\mathbf{q}$.
- Our method using HITS ($\mathbf{QA}_{hits}$):
  This method outputs the top 15 adjective facets that are ranked highest by Equation (1). We used Yahoo! Japan Chiebukuro’s API and obtained 1,000 questions and related answers.
- Our method using HITS and query association ($\mathbf{QA}_{h+q}$):
  This method outputs the top 15 adjective facets that are ranked highest by Equation (2). We set parameter $\lambda_f = 0.5$.

3.2 Method

To evaluate the effectiveness of our method with diverse topics, we prepared 10 categories and six queries for each category (60 queries in total). The categories and example queries used in the experiment are listed in Table 1.

Table 2: Average scores (AVG) and MAP@k for the four methods (highest values in bold).

	AVG	MAP@5	MAP@10	MAP@15
QS	2.699	0.561	0.502	0.478
WEB	2.612	0.532	0.498	0.483
QA$_{hits}$	2.508	0.507	0.489	0.469
QA$_{h+q}$	**2.701**	**0.592**	**0.566**	**0.536**

Table 3: Average scores for each category (highest values in bold).

Category	QS	WEB	QA$_{hits}$	QA$_{h+q}$
Sports	2.492	2.210	2.813	**3.163**
Science&Academics	**2.472**	2.167	2.032	2.290
Politics&News	**2.421**	2.393	2.194	2.349
Travel	2.881	2.849	**3.536**	3.440
PC&Electronics	**3.032**	2.530	2.478	2.534
Health	**2.992**	2.810	2.571	2.639
Business&Finance	**3.004**	2.905	2.429	2.627
Relationship&Life	2.758	**3.024**	2.262	2.488
Education&Parenting	2.397	2.437	2.159	**2.452**
Home&Food	2.544	2.802	2.603	**3.028**

We asked six volunteers to participate in the experiment. We divided the 60 queries into two groups and each participant evaluated 30 queries (10 categories × 3 queries.) The order of the queries showed to each volunteers were balanced. The experiment proceeded as follows.

- We first gave short descriptions of the required information (e.g., "You are planning to travel to Kyoto" or "You are planning to purchase a new smartphone").
- Next, we showed the participants 60 terms, which had been extracted from the four methods described in Section 3.1. The 60 terms were randomly placed.
- For each term, participants were asked to judge how much they wanted to check the information suggested for it in a five-point Likert scale (1 = *not interested at all* and 5 = *strongly interested*).

3.3 Results

Table 2 shows the results for average score (AVG) and mean average precision at cutoff $k = 5, 10$ and 15 (MAP@k) of the four methods. Terms that scored 4 and 5 were deemed relevant and terms that scored 1, 2, and 3 were deemed irrelevant to measure mean average precision. From the results in Table 2, we can see that our proposed method **QA**$_{h+q}$ outperformed **WEB** and **QA**$_{hits}$ in terms of all measures. We also found that the average scores of **QS** and **QA**$_{h+q}$ were the almost same. simply considering the relationships between adjective facets and co-occurring entities. This results indicate that adjective facets extracted from Q&A corpus can attract users' interests and that a Q&A corpus can be a valuable resource to support exploratory Web searches.

We further analyzed the experimental results in terms of the categories used in the experiment. Table 3 shows the average scores of the four methods for each category. We can see that the effectiveness of our method depended on the category. For example, **QA**$_{h+q}$ got over 3.00 average score in the "Sports", "Travel", and "Home & Food" categories. On the other hand, it got much lower average score in the "Science & Academic" and "Politics & News" categories. From this result, we can estimate that the importance of adjectives depend on the types of knowledge that users are required. When users are searching for information related to a topic that requires highly technical or specialized knowledge of them, objective opinions or technical terms might be more important to inform their information needs of a Web search engine than adjectives. With such topics, like politics or science, adjective facets might not preferred by users since those topics require advanced knowledge of users. On the other hand, in the topics related to the "Sports" and "Travel", "Home & Food" categories, subjective opinions or viewpoints might be important to find information, thus, adjective facets were preferred by the participants.

Table 4 shows example adjective facets that our method **QA**$_{h+q}$ suggested to the participants. Since these terms are seldom suggested by conventional query suggestions, our method has the potential of complementing the conventional query suggestions.

Table 4: Example adjective facets.

Query	Examples	
Japanese	great hitters young players	favorite teams cool players
Kyoto travel	famous temples beautiful spots	minor places cheap hotels
Influenza	severe headache cold chill	accurate knowledge correct hand-washing

4. CONCLUSIONS

Our experimental results showed that our approach could suggest more terms that users judged interesting than other baselines in several topics. We plan to further analyze the relationships between the topics and effectiveness of adjective facets. Moreover, we plan to develop a ranking algorithm for queries that contain adjective facets and conduct experiments to evaluate the effectiveness our adjective facets in actual search scenarios.

5. ACKNOWLEDGMENTS

This work was supported in part by Grant-in-Aid for JSPS Fellows (#09J55302), Grant-in-Aid for challenging Exploratory Research (#22650018), Grant-in-Aid for Young Scientists (A) (#23680006), and by "Informatics Education and Research Center for Knowledge-Circulating Society" (Project Leader: Katsumi Tanaka, MEXT Global COE Program, Kyoto University).

6. REFERENCES

[1] B. Croft, D. Metzler, and T. Strohman. *Search Engines: Information Retrieval in Practice*. Addison-Wesley Publishing Company, USA, 2009.

[2] H. Cui, J. Wen, J. Nie, and W. Ma. Probabilistic Query Expansion Using Query Logs. In *Proceedings of the 11th International Conference on World Wide Web*, pages 325–332, 2002.

[3] H. Deng, M. R. Lyu, and I. King. A Generalized Co-HITS Algorithm and its Application to Bipartite Graphs. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 239–248, 2009.

[4] J. M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. *J. ACM*, 46:604–632, 1999.

[5] R. White and R. Roth. Exploratory Search: Beyond the Query-Response Paradigm. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 1(1):1–98, 2009.