Looking for Non-existent Information: A Consumer-led Interactive Search Approach

Yuqing Mao^{1,2}

Haifeng Shen²

Chengzheng Sun¹

¹ School of Computer Engineering, Nanyang Technological University, Singapore

² School of Computer Science, Engineering & Mathematics, Flinders University, Australia
{mao.yuqing, haifeng.shen}@flinders.edu.au, czsun@ntu.edu.sg

Exponential growth of user-generated content (UGC) on the Web demands an effective way for information consumers to acquire what they need. Specialized UGC search services outperform general-purpose search engines that usually have difficulties retrieving UGC or rank them low. However, with existing UGC search services, consumers may only be able to acquire information that is related to their needs rather than the exact information what they want. In this paper, we present a novel consumer-led interactive search approach that can help consumers acquire the information they exactly want through a consumer-led interactive search process where invited information providers jointly create such information on the fly. This approach has been implemented in a prototype system, which uses an epistemology structure to represent a consumer's information needs and lead an interactive search process. Preliminary user feedback has shown that this approach is particularly effective for a consumer to acquire a structured knowledge unit consisting of diverse but coherent information.

User-generated content, social media search, epistemology.

1. INTRODUCTION

With the popularity of social media, user-generated content (UGC), especially content generated in social networking and micro-blogging sites, has greatly increased on Internet in recent years, because publication and circulation of information has become extremely easy and rapid. For instance, currently there are more than 230 million pieces of real-time UGC (e.g. *Facebook's* newsfeed or *Twitter's* tweet) published each day and 40% of all searches are estimated to have some sort of demand for such information (Spark, 2009).

General-purpose search engines (e.g. *Google* or *Bing*) are unable to follow after UGC well because the crawlers need to take significantly long time to discover and index each new page. Moreover, they are unable to properly display UGC to the users because the content would not be displayed on the top of the search results (due to low PageRank).

Specialized UGC search services, such as *Twitter Search* or *Google Real-time Search*, partially solve general-purpose search engines' problems with UGC by using user-provided keywords to constantly keep up with UGC updates on a certain topic/person, e.g. activities, blog posts, newsfeeds, and tweets, in social media.

Nevertheless, these search services are still similar to general-purpose search engines that locate resources on the Web according to users' queries. If a query does not retrieve the information that a consumer (i.e. an information seeker) exactly wants, she/he has to amend the query using different keywords. If such an iterative trial-and-error process does not end up with what she/he wants, the information is deemed non-existent and she/he would give up and try again some time later.

However, content generated by ordinary users rather than professional Web site editors implies that the information providers could prepare or update their content to meet the needs of the information consumers. Especially for the content generated by numerous users in the popular social media platforms such as *Twitter* and *Facebook*, if the providers who have published some information are informed of what else consumers are looking for, they would be able to generate such content. For example, in the study of Teevan et al (Teevan et.al. 2011), a lot of people published tweets about the movie *New Moon*. However, there are many questions about whether the movie was worth seeing. If the provider who published content related to "new moon" were informed of the queries such as "recommendation", she/he would publish a new tweet, e.g., "I recommend the movie new moon that is as good as the movie twilight".

In this paper, we present a novel consumer-led interactive search approach called *CLEMIS* (Consumer-Led Epistemology-Mediated Interactive Search), which can help consumers acquire the non-existent information they exactly want through a consumer-led interactive search process.

In the *CLEMIS* approach, an information consumer leads an interactive search process through a pre-structured epistemology (Mao et.al. 2010) representing her/his personalized information needs on a specialized UGC search system, which retrieves related UGC according to the epistemology and then informs the providers who published the content of this epistemology because they are likely to have the knowledge to publish the exact information the consumer is looking for. The consumer can interact with the providers to clarify, comment on, refine, or request more information on the infused epistemology.

The *CLEMIS* approach has been implemented in a prototype system. For the purpose of conducting internal usability testing, we have also implemented a micro-blogging system for information providers to publish UGC. Preliminary user feedback has shown that this approach is particularly effective for a consumer to acquire a structured knowledge unit consisting of diverse but coherent information.

The rest of this paper is organized as follows. First we describe some work related to UGC search. Then we sketch the *CLEMIS* approach to tackling the address issue. After that, we present a prototype consumer-led interactive search system that implements the proposed approach, followed by discussions of a preliminary usability testing. Finally we conclude the paper with a summary of major contributions and future work.

2. RELATED WORK

Recent years have witnessed a phenomenal growth of user-generated content on the Web, mainly owing to the increasing popularity of various micro-blogging systems and social networking sites. For example, *Twitter* users can publish real-time topical news (Phelan et.al. 2009). However, without a proper search engine, the vast majority of such content is only visible to certain social networking contacts rather than reaching general public.

UGC search is therefore a new compelling area that has dramatically changed the traditional information acquisition model. For example, users can search for latest tweets posted on *Twitter*, or friends' newsfeed on *Facebook*, with their built-in search services. However, these UGC search services mostly do nothing more than simply

collecting user-generated news and then displaying them to the user in the chronological order. Such user experience is insufficient in UGC search because people were particularly likely to search for UGC on a topic of interest. For example, according to the investigation on the user behavior in micro-blogging search (Jansen et.al. 2010), lots of users show long-term interest in a certain topic (more than just a news item), as a variety of inter-related queries are often repeated over time.

It is worth discussing some unconventional search systems that also exhibit some relevance to the consumer-led process, although they are not specially catered for UGC search. For example, Q&A (Question and Answer) systems such as *Yahoo! Answers* and *Baidu Zhidao* can be regarded as a consumer-led process as a user directly posts her/his question on the system in order to lead provider(s) to prepare their answers accordingly. Some personalized or vertical search systems based on interest-based ranking (Xu et.al. 2008) or topic-specified crawling (Menczer et.al. 2004) can also be regarded as a consumer-led process as a user's profile (instead of a topic query) is used to lead the search. Some social search systems can also acquire information from the providers based on the social interaction in search processes (Fu 2008, Kamber et.al. 2009). For example, *Aardvark* (Horowitz and Kamvar 2010) provides various communication tools (e.g. instant message, email, etc.) for a user to interact with her/his friends (information providers) during a search process.

Although different work was essentially associated with a consumer-led process largely due to its specific problem context, the *CLEMIS* approach support the consumer-led interactive search in a very unique problem context of meeting individual consumer's personalized and diverse information needs for UGC. That is, a consumer could acquire the information she/he exactly wants through a consumer-led interactive search process where invited information providers jointly contribute such information to the consumer-defined well-structured knowledge unit regarding a specific topic.

3. THE CLEMIS APPROACH

3.1 Overview

A schematic architecture of the *CLEMIS* approach is shown in Figure 1(a), which sketches key components that work together to support consumer-led interactive search by joint construction of the pre-structured epistemology.

Epistemology Constructor – for a consumer to create a pre-structured epistemology depicting a

blueprint to lead an interactive search process. The structure of epistemology will be discussed shortly.

Epistemology Index - for the *Filter* component to quickly retrieve the epistemology that can be filled out with relevant UGC.

User-Generated Content Cache – for the *Filter* component to discover cached UGC from external social media systems that is related to an epistemology created by the consumer.

Filter – a core component that retrieves *User-Generated Content Cache* and *Epistemology Index* in order to find matches between relevant fields of an epistemology and cached UGC. In addition, it collects the contact information of the providers whose content matches the epistemology so that the *Dispatcher* component can dispatch the epistemology to them in order to lead the search process and facilitate interaction between the consumer and these providers via the *Communicator* component. It also passes on the matched UGC from the cache to the *Writer* component, which will then fill them into the relevant fields of the epistemology.

Mediator – the bridge between the consumer-led interactive search system and external user-generated resources provided by micro-blogging users or social networking friends.

3.2 The Pre-structured Epistemology

Epistemology is structured hierarchically, as shown in Figure 1(b). An epistemology, which describes a consumer's information needs for a specific topic, consists of a list of separate but inter-related *fields* (a.k.a. sub-topics) and each field is composed of a set of independent or inter-related *threads* (for interaction between consumer and providers).

Field is the working unit in an epistemology. Each field is tagged with a set of sensible keywords that will be used by the *Filter* component to precisely match with cached UGC. Matched content will be filled into the field by the *Writer* component. *Thread* is the interaction and cooperation unit. The consumer can actively interact with information provider(s) that have contributed to a specific field by commenting on their input in order to clarify doubts, correct errors, or polish the results. Multiple information providers may jointly input their content to the different threads of the same field, or even different threads of different fields, no matter whether they are aware of their cooperation.

The epistemology structure is obviously advantageous in the case that a consumer wants to search for updates on a topic that cannot be simply described by a few keywords. With an ordinary

UGC search services, the consumer has to generate multiple queries using different keywords on multiple instances of a Web browser because using all keywords in a single query is likely to get no matched result at all.

In contrast, with the consumer-led interactive search system using structured epistemology, multiple queries using different keywords can be generated simultaneously and matched results will be filled into the corresponding fields of the epistemology automatically and simultaneously. More importantly, generation of multiple queries is completely transparent to the consumer, who only sees what she/he wants to know have been filled with results and who may wish to take the chance to interact with the information providers just for the sake of polishing the results. Furthermore, the consumer can always interact with providers to clarify doubts, correct errors, or polish the results.

4. THE CONSUMER-LED INTERACTIVE SEARCH SYSTEM

We have applied the *CLEMIS* approach to the design and implementation of the consumer-led interactive search service in the *Baijia* prototype system. We have also implemented a micro-blogging system for information providers to publish content for the purpose of conducting internal usability testing. We will discuss some user interface features of the systems in this section. Note that the external micro-blogging system is connected to the search system via the *Mediator* component within the system by using the micro-blogging system's API. Therefore it is also possible to connect the search system with various external social media system via their APIs, e.g. *Twitter*, *Facebook*, and so on.

4.1 Epistemology Constructor and Filter Interface

In our system, a consumer can create a pre-structured epistemology and use queries and related phases to present the information needs as far as possible. As shown in Figure 1(c), the consumer is seeking for information about the "World Cup Final". She/he would like to acquire information about both sides of the match, therefore a pre-structured epistemology is created with two fields: "Spain in World Cup", and "Netherlands in World Cup". Each field will be a container for related UGC.

In the first field, the consumer is interested in the activities of Spain in the World Cup. Therefore she/he can define some rules to filter the UGC that meets the information needs. For example: "Expect these words" such as "coach", "the first eleven", and so on, in the user-generated content. The

search service will generate the conditional expression “include (Spain in World Cup) AND (coach OR (the first eleven))” in the epistemology index. Once there is information in the user-generated content cache that satisfies the condition, e.g. “Spain coach Del Bosque ...”, it will be inserted into the field immediately. At the same time, the inter-related phase “the first eleven” will be dispatched to suggest the providers to publish relevant content.

The consumer can also filter the information with specified providers, e.g. the content from “Messi”; or filter according to the expertise levels of providers, e.g. grade, ranking by viewers, the number of the followers; or filter according to time, e.g. “first publishing” or “last updating”.

4.2 Micro-blogging Interface

Figure 1(d) shows the interface of the micro-blogging system connected with the consumer-led interactive search system. Providers can easily publish content through such a *Twitter*-like interface.

While a provider types a piece of content, relevant epistemologies dispatched from the consumer-led interactive search system are being filtered in and displayed so as to suggest her/him to give as much relevant information as needed by consumers.

For example, while a provider types “Spain sweat over Villa injury”, the micro-blogging system immediately sends it to the consumer-led interactive search system, which in turn filters all pre-structured epistemologies to find those asking for such information, e.g. the epistemologies titled “Spain World Cup” and “Expect these words: injury, influence”.

The consumer-led interactive search system then dispatches these matched epistemologies to the micro-blogging system and posts a micro-blog there, e.g. “Currently interested information about #Spain #injury: #substitution(12) #squad(5) #influence(1)”. Such information will suggest the provider to publish the content that is most interested (i.e. many consumers are interested in the substitution for Villa) based on the number following each tag indicates, which is calculated as:

$$N = N_{epi} - N_{rec}$$

where N_{epi} is the number of epistemologies are expecting content tagged by that keyword, and N_{rec} is the number of providers who has published any content tagged by that keyword.

4.3 Writer and Communicator Interface

At the same time, the writer component will update the pre-structured epistemologies that are requiring such information by filling a user-defined field with the relevant content identified by the Filter

Figure 1: (a) The schematic architecture of CLEMIS; (b) The structure of the epistemology; (c) The epistemology constructor and filter interface; (d) The micro-blogging interface; (e) The writer and communicator interface with the epistemology

component, as shown in Figure 1(e). Relevant content is identified using the traditional approach in information retrieval (IR) research: term frequency – inverse document frequency (TF-IDF) (Salton and Buckley, 1988).

For each term t_i of the query in the user-defined field, $tf(t_i)$ is the frequency of t_i in UGC, and we calculate the IDF in the form:

$$idf(t_i) = \log \frac{N}{df(t_i)}$$

where N is the number of all UGC and $df(t_i)$ is number of UGC where the term t_i appears.

Then: $tf-idf(t_i) = tf(t_i) \times idf(t_i)$

However, for some UGC such as tweet, since users tend to remove word redundancy from a tweet to save space, seldom terms are repeated in a tweet. In that case, TF-IDF is essentially just the IDF term.

The content and comments are organized based on different providers. Therefore those providers can work jointly and concurrently while the structure of the epistemology remains unchanged. That could make the epistemology more readable than search results ordered by time.

Thereafter, the consumer is able to interact with the provider(s). For example, the consumer can comment on the thread “*Pele: Spain Are Favourites to Win World Cup*” by “*How about the prediction of Octopus Paul*” and the provider can publish another piece of content to update the epistemology accordingly.

5. USER FEEDBACK

We conducted a preliminary usability testing of the consumer-led interactive search system in order to: 1) understand whether users (i.e. information consumers and providers) like this new concept and if yes what features they particularly like, 2) investigate what kind of search tasks where the consumer-led interactive search system does better than existing UGC search services, 3) study whether the system is easy-to-use and what special skills users need to use the system, and 4) get some feedback to improve the system.

5.1 Participants and Tasks

Ten users participated in this small-scale user research. The users included both men and women, and ages ranged from around 20 to around 50 years old (median=27). Five of them were undergraduate and postgraduate students, three were staffs, and the rest were IT professionals. As we chose *Twitter* as a representative UGC media, eight of them have the experience of using UGC search services and seven of them are current

users of *Twitter* and have the experience of publishing content on *Twitter*.

To investigate what kind of search tasks where the consumer-led interactive search system might do better than existing search systems, we deliberately designed two search tasks.

The first scenario was searching for UGC about the *new generation iPhone*. Since the consumers were eager to know more details about that product from a current owner or an expert, it was very likely for the consumer to invite the providers to publish their diverse views to the consumer’s information needs and such personal knowledge could not be retrieved through existing search systems.

The second scenario was searching for UGC about the *World Cup Final*. Since the consumers were pretty clear about what they were looking for, e.g. goals, shoots, etc., they could clearly define an epistemology structure to invite the providers to publish content that meets their needs.

The same set of participants did the two scenarios first with *Twitter Search* and with the UGC published on *Twitter* and then with the *Baijia* consumer-led interactive search system and with the UGC published on our micro-blogging system. They were not allowed to use any other communication channels, e.g. phones or instant messengers, except the systems given to them.

At the end of the testing, we interviewed the participants in order to understand their views on the system, in terms of novelty and search results quality, as well as the advantages and limitations of the system.

5.2 Scenario 1

In this scenario, seven consumers were interested in buying the new generation iPhone. However, they have been bored by the perpetual advertisements and stereotyped reviews on the Web. Therefore, they turned to search for UGC about the latest and just critiques on the product. Three providers were either the current owners or technical experts of the product.

Participants were impressed by the *Baijia*’s ability of allowing consumers to lead the interactive search process, and invite multiple providers to publish their content incorporated in different fields of the epistemology in the search process. Since the consumers would make the decision for buying the product based on the search results, they had to search for advantages and disadvantages reflected by the owners. Most of such information could not be retrieved by *Twitter Search* since many providers generally would not publish their

personal views on specific details. *Baijia* was able to suggest them to publish required information through promoting with epistemologies created by consumers on the same topic as the content already generated by the provider, such as “my latest iPhone”.

It is because the search process is led by the consumer that makes the epistemologies so valuable to a particular consumer for her/his particular information needs. All participants commented on that feature during the interview.

5.3 Scenario 2

In this scenario, the participants were situated in context of the 2010 World Cup final. Six participants were designated as consumers to search what they were interested in the match, while the other four participants were designated as providers to publish content.

All participants agreed that the consumer-led interactive search system brought a novel user experience, both for consumers and providers. As the search was mediated by structured epistemologies, it was particularly efficient for consumers to complete a search task that generates multiple or complex queries; it was also effective for providers to generate high quality content that could meet the more consumers' needs. Further, the quality of search results of the *Baijia* search system was much more satisfied than *Twitter Search*, not only because they could get more information through interaction with the provider, but also because it was really hard to read on some topic in *Twitter Search* which only had a simple single column filled with feed. In contrast, in the structured epistemology created by the consumer, the UGC was catalogued into fields with multiple queries simultaneously and automatically, which was more readable than a sequence of content that updated frequently.

6. CONCLUSION AND FUTURE WORK

Recent years have witnessed a phenomenal growth of UGC on the Web, mainly owing to the increasing popularity of various micro-blogging systems and social networking sites. As opposed to the immense amount of UGC available in the cyberworld, tools for effectively helping get such content across to general public (instead of within certain social networking contacts) are lacking. General-purpose search engines and specialized UGC search services could not help consumers acquire the exact information they want. The proposed *CLEMIS* approach can help consumers acquire the information they exactly want through a consumer-led interactive search process where

invited information providers jointly create such information on the fly.

Furthermore, we have implemented the approach in the *Baijia* prototype system, which uses an epistemology structure to represent a consumer's information needs and also lead an interactive search process. Initial usability testing of the system has given positive feedback to the approach, but we are conscious that we need to conduct more rigorous evaluation tasks ahead.

We are improving the prototype system based on the initial feedback. We are working towards connecting our system with public social media systems, e.g. getting UGC through “*Twitter Firehose*” so that we can conduct more realistic and rigorous evaluation tasks.

7. REFERENCES

- Fu, W. (2008) The microstructures of social tagging: a rational model. In *Proc. CSCW'08*, ACM Press, 229-238.
- Horowitz, D. and Kamvar, S. D. (2010) The anatomy of a large-scale social search engine. In *Proc. WWW '10*. ACM Press, 431-440.
- Jansen, B. J., Campbell, G., and Gregg, M. (2010) Real time search user behavior. *Ext. Abstracts CHI 2010*. ACM Press, 3961-3966.
- Kamerer, Y., Nairn, R., Piroli, P., and Chi, E. H. (2009) Signpost from the masses: learning effects in an exploratory social tag search browser. In *Proc. CHI'09*, ACM Press, 625-634.
- Mao, Y., Shen, H., and Sun, C. (2010) Supporting exploratory information seeking by epistemology-based social search. In *Proc. IUI '10*. ACM Press, 353-356.
- Menczer, F., Pant, G., and Srinivasan, P. (2004) Topical web crawlers: Evaluating adaptive algorithms. *ACM Trans. Internet Technol.* 4(4), 378-419.
- Phelan, O., McCarthy, K., and Smyth, B. (2009) Using twitter to recommend real-time topical news. In *Proc. RecSys '09*. ACM Press, 385-388.
- Salton, G. and Buckley, C. (1988) Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5), 513-523.
- Spark, D. (2009) *Real-Time Search and Discovery of the Social Web*. Spark Media Solutions. <http://www.sparkminute.com/?p=1261>
- Teevan, J., Ramage, D., and Morris, M. R. (2011) #TwitterSearch: a comparison of microblog search and web search. In *Proc. WSDM '11*. ACM Press, 35-44.
- Xu, S., Bao, S., Fei, B., Su, Z., and Yu, Y. (2008) Exploring folksonomy for personalized search. In *Proc. SIGIR '08*. ACM Press, 155-162.