CONSENTO: A Consensus Search Engine for Answering Subjective Queries

Jaehoon Choi Donghyeon Kim Seongsoon Kim Junkyu Lee
Sangrak Lim Sunwon Lee Jaewoo Kang
Korea University
Seoul, Korea
{ jaehoon, donghyeon, seongkim, onleejk, roghdejd, sunwonl, kangj }@korea.ac.kr

ABSTRACT

Search engines have become an important decision making tool today. Decision making queries are often subjective, such as "best sedan for family use," "best action movies in 2010," to name a few. Unfortunately, such queries cannot be answered properly by conventional search systems. In order to address this problem, we introduce CONSENTO, a consensus search engine designed to answer subjective queries. CONSENTO performs subdocument-level indexing to more precisely capture semantics from user opinions. We also introduce a new ranking method, or ConsensusRank that counts in online comments referring to an entity as a weighted vote to the entity. We validated the framework with an empirical study using the data on movie reviews.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Retrieval models

General Terms

Design, Algorithms, Experimentation

Keywords

entity search, consensus search, ConsensusRank, maximal coherent semantic unit, segment indexing

1. INTRODUCTION

Web search has become ubiquitous today. Commercial search engines have been highly effective for factual queries such as "iPhone 4S release date." Users can find the answers to such queries in one of the top ranked documents. However, the current search engines fall short of giving proper answers to subjective queries. For example, queries like "best action movies in 2010" and "thrillers with plot twist" are not properly answered by the current engines. There, the top ranked documents may contain a list of best action movies or plot-twisting thrillers. That list, however, reflects only document authors' opinions, rather than the public sentiment.

Apart from document retrieval systems, there exist entity search and question-answering (QA) systems. These systems produce direct answers to queries such as "romantic comedies in 2011" and "airlines flying boeing 747"[1] [3, 1]. Still, they just focus on factual queries that contain finite sets of true answers. Thus, these systems are not appropriate, either, to our problem context.

We term the problem we are to address as the *consensus search problem*, which can be defined as follows: Given an entity set $E = \{e_1, e_2, \cdots, e_u\}$ and a query q, suppose there exists an ideal ranking function $CR(q, e_i)$ that would return a rank of e_i reflecting the amount of votes that e_i would have received from a long-running online poll on query q. The consensus search problem is then defined as the problem of finding the ranked list of entities $L = [e_{k_1}, e_{k_2}, \cdots, e_{k_u}]$ such that $CR(q, e_{k_i}) \geq CR(q, e_{k_j})$ for all $1 \leq i \leq j \leq u$.

Although consensus search is yet to be used widely, such consensus queries are assuming more importance as web search has become an essential tool in diverse decision making scenarios, such as online shopping, political sentiment analysis, and business intelligence. Moreover, also on the rapid rise are consumer reviews and comments available on the Web and social media, naturally incurring demands for new search services mainly designed to process the social data. In view of these problems, we introduce CONSENTO, a consensus search engine.

2. CONSENTO ARCHITECTURE

Figure 1 illustrates the architecture of CONSENTO. CONSENTO is built on an open source search engine with minimal modifications to the logical structure of indexes and the ranking scheme. This ensures the scalability of conventional search engines. CONSENTO consists of two subsystems: the indexing subsystem (①②③ in Figure 1) and the searching subsystem (④⑤⑥ in Figure 1). The indexing subsystem is essentially identical to conventional document-indexing systems except for the fact that it performs subdocument-level indexing.

Unlike conventional search engines, CONSENTO indexes Maximal Coherent Semantic Unit (MCSU), which is a maximal subsequence of words containing a single coherent semantic within a document. For example, "excellent performance, but plot was hard to follow" implies two different sentiments, or one positive semtiment (performance) and the other negative (plot). However, for a query "excellent plot," conventional text retrieval systems would find the above review relevant to the query, since their indexing unit is a document and the review document contains both of the two query terms in proximity. In order to capture the user's

[1]TREC 2011 Entity. http://ilps.science.uva.nl/trec-entity/

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

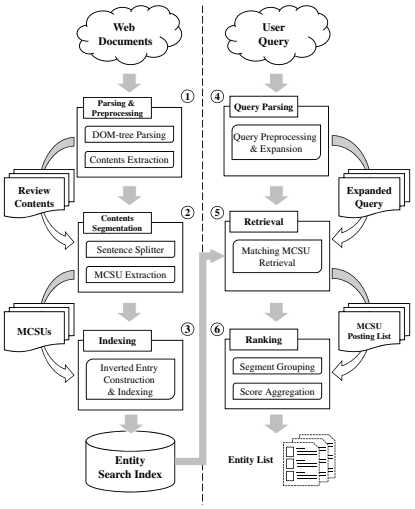

Figure 1: CONSENTO architecture.

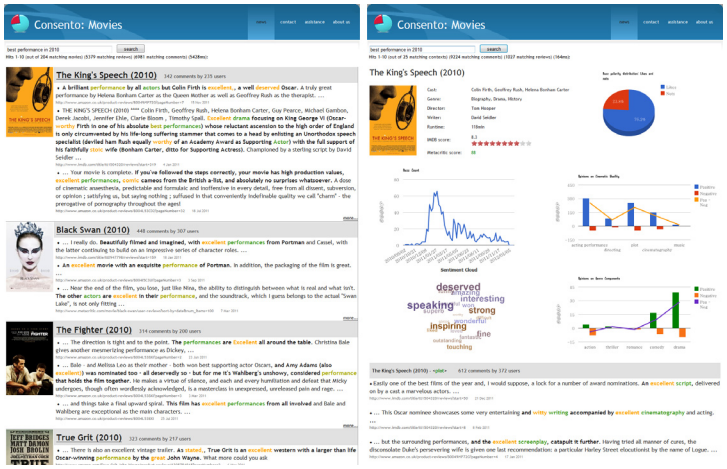

Figure 2: Consento results page for query "best performance in 2010" (left) and its details page for movie "The King's Speech" (right)

Table 1: nDCG@10 results with 2010 movies.

Query (aspects)	VSM+BM base	VSM+BM OE(+QAM)	BM25 base	BM25 OE(+QAM)	CR(gain)
single	0.19	**0.34**(n/a)	0.17	0.27(n/a)	**0.60**(**76.5%**)
multi	0.20	**0.39**(0.38)	0.17	0.33(0.38)	**0.63**(**61.5%**)

sentiment correctly, CONSENTO splits the review comment into two MCSU segments, and indexes them separately (② in Figure 1).

Among others, our searching subsystem significantlly differs from conventional standard text retieval models in terms of ranking. Conventional models produce segments that best match query terms. On the other hand, CONSENTO is designed to return entities that are most agreed upon by users with respect to the query context. In order to implement this, we introduce a new ranking algorithm, ConsensusRank, which counts an online comment matching a particular query as a weighted vote to the "referred to" entity (⑥ in Figure 1).

In particular, given a query, all matching segments are retrieved from the index. The retrieved segments are then grouped by their referencing entities. Finally, the scores of the segments are aggregated to compute the scores of the corresponding entities. The score of each segment is determined by multiple factors including similarity to the query terms, sentiment orientation, review quality, source authority, and recency. Figure 2 shows the examples of the CONSENTO results pages.

3. VALIDATION

For validation, we used movie reviews crawled from six major sources including IMDB, Amazon, Metacritic, Flixster, Rottentomatoes and Yahoo Movies. More than 890K reviews were collected for more than 10K movies. The current version of CONSENTO is built on Apache Lucene 3.1. As no gold standard is available for consensus movie ranking, we constructed one for validation purposes, using the award histories of top 12 award ceremonies and festivals such as the Academy Awards and the Cannes Film Festival. The winner in an event receives 2 points and a nominee receives 1 point. By aggregating the points, we generated the movie rankings for five award categories each year. The categories include "best movies" (e.g., best picture award), "best performance," "best directing," "best music," and "best screenplay."

As the baseline, we used Ganesan and Zhai's Opinion Expansion(OE) and Query Aspect Modeling(QAM) approaches [2], which works as follows. They concatenate all reviews for a film in one document, and indexes the documents using a standard text retrieval system. As to query time, OE expands the user query with a predefined set of opinion word synonyms, and processes the expanded query as usual. QAM is an additional improvement that splits a query based on aspects, processes each subquery sperately, and aggregates the scores from the subqueries to compute final rankings. Finally, the ranks of the returned documents are the ranks of the corresponding entities. It was reported that the method is simple and yet effective for opinion-based entity ranking.

Table 1 shows the nDCG@10 scores for single and multi-aspect(2-5 aspects) queries on 2010 movies. As shown, CONSENTO(CR) outperformed the baselines, VSM+BM(Lucene defualt), BM25, and their OE and QAM extensions, with significant margins. The working prototype of CONSENTO is available at http://CONSENTO.korea.ac.kr.

4. ACKNOWLEDGMENTS

This project was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MEST) (2009-0077688, 2009-0086140).

5. REFERENCES

[1] S. Chakrabarti, D. Sane, and G. Ramakrishnan. Web-scale entity-relation search architecture. In *WWW '11*, pages 21–22, 2011.

[2] K. Ganesan and C. Zhai. Opinion-based entity ranking. *Information Retrieval*, 2011.

[3] J. Lin and B. Katz. Question answering from the web using knowledge annotation and knowledge mining techniques. In *CIKM '03*, pages 116–123, 2003.