Smith Search: Opinion-Based Restaurant Search Engine

Jaehoon Choi
Opinion8, Inc.
Seoul, Korea
jchoi@opinion8.io

Donghyeon Kim
Opinion8, Inc.
Seoul, Korea
dkim@opinion8.io

Donghee Choi
Opinion8, Inc.
Seoul, Korea
dchoi@opinion8.io

Sangrak Lim
Korea University
Seoul, Korea
limsangrak@korea.ac.kr

Seongsoon Kim
Korea University
Seoul, Korea
seongkim@korea.ac.kr

Jaewoo Kang[*]
Korea University
Seoul, Korea
kangj@korea.ac.kr

ABSTRACT

Search engines have become an important decision-making tool today. Unfortunately, they still need to improve in answering complex queries. The answers to complex decision-making queries such as “best burgers and fries” and “good restaurants for anniversary dinner,” are often subjective. The most relevant answer to the query can be obtained by only collecting people’s opinions about the query, which are expressed in various venues on the Web. Collected opinions are converted into a “consensus” list. All of this should be processed at query time, which is impossible under the current search paradigm. To address this problem, we introduce Smith, a novel opinion-based restaurant search engine. Smith actively processes opinions on the Web, blogs, review boards, and other forms of social media at index time, and produces consensus answers from opinions at query time. The Smith search app (iOS) is available for download at http://www.smithsearches.com/introduction/.

Categories and Subject Descriptors: H.3.1 [Content Analysis and Indexing]: Indexing methods, Linguistic processing H.3.3 [Information Storage and Retrieval]: Retrieval models

General Terms: Design, Algorithms, Experimentation

1. INTRODUCTION

Web search is commonly involved in the decision-making process. However, answers to decision-making queries are often subjective and the current search engines fail to deliver satisfactory results. Current search engines are effective in providing small numbers (typically one) of correct answers to “fact-finding” queries. For the query “address of Umami Burger in San Francisco,” most current search engines place the restaurant’s website at the top of the results page. However, the query “best burgers and fries in San Francisco” is more complex because the most useful answer to the query may be found in more than one document. A document may list the best burger restaurants but it reflects only the personal preferences of the person who wrote the document and not the consensus of the public.

Perhaps the most reliable approach to answering subjective questions is to ask as many people as possible for their opinions. However, this approach is often infeasible because an individual’s social network may be too small. And even if one has a large enough network, there may be an insufficient number of people who can offer their opinions.

Another approach involves processing online comments posted by other users. For many thinkable questions, chances are that people already have expressed their opinions somewhere on the Web, blogs, review boards, and other forms of social media. Users such as restaurant-goers and shoppers read as many posts as possible before making their final decision. The reliability of their decision improves as the number of reviews and comments they process increases. However, this is a very time-consuming and labor-intensive process.

To address the problem particularly in the Restaurant domain, we introduce Smith, our opinion-based restaurant search engine. Smith is built on Consento, a vertical entity search engine, which we introduced in [1]. In [1], we validated its efficacy for Movie and Hotel search. In this work, we aim to demonstrate its ability to expand domains applying it to the new Restaurant domain.

Smith actively processes people’s opinions on the Web at index time, and produces a “consensus” list for any ad-hoc query at query time using the index. Smith employs the following two novel methods. First, Smith indexes logical entities (such as Burger Joint and Shake Shack) rather than physical documents. Smith divides documents into short passages each of which may contain a user’s opinion on an aspect of a restaurant. The restaurant is then indexed along with the passage that describes it.

Second, Smith takes a unique ranking approach that is significantly different from conventional search methods. While conventional systems return documents that are the most relevant to the terms of a query, Smith returns the most agreed upon *entities* in reviews and comments on the Web with respect to the query context. To implement this, we introduce a new ranking model ConsensusRank which considers a user’s opinion that matches a particular query as a weighted vote for the “referred to” entity. In particular, given a query, all matching passages are retrieved from the index.

[*]Corresponding author.

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742829.

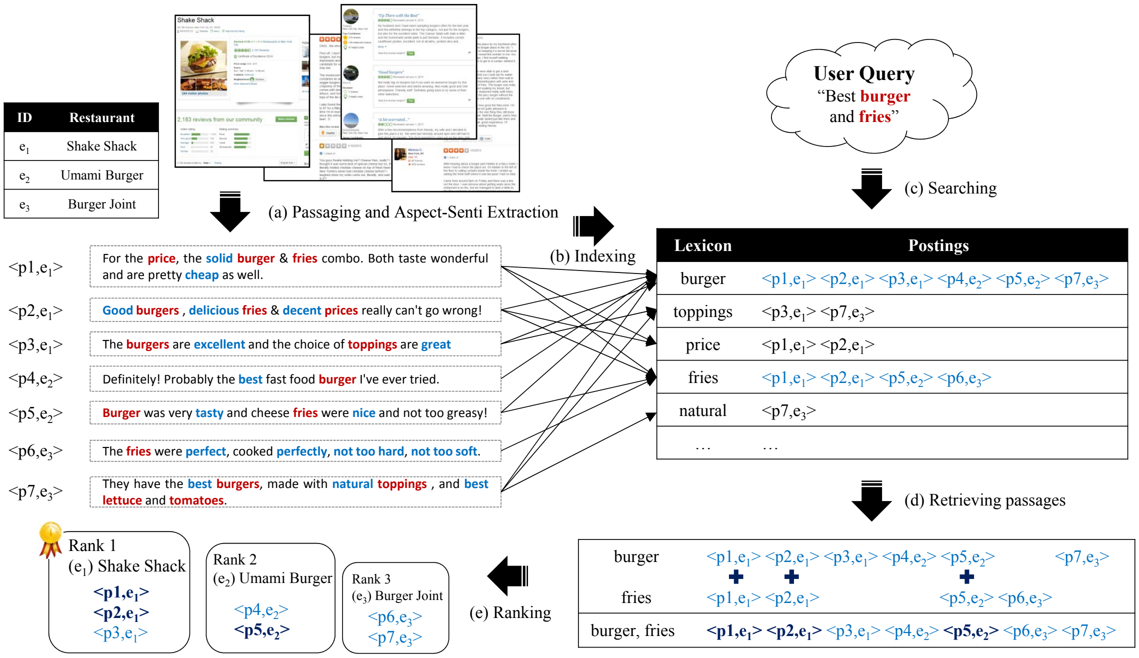

ID
Restaurant
e_1 Shake Shack
e_2 Umami Burger
e_3 Burger Joint
User Query
"Best burger and fries"
(a) Passaging and Aspect-Senti Extraction
(b) Indexing
(c) Searching
(d) Retrieving passages
(e) Ranking
<p1,e_1> For the price, the solid burger & fries combo. Both taste wonderful and are pretty cheap as well.
<p2,e_1> Good burgers , delicious fries & decent prices really can't go wrong!
<p3,e_1> The burgers are excellent and the choice of toppings are great
<p4,e_2> Definitely! Probably the best fast food burger I've ever tried.
<p5,e_2> Burger was very tasty and cheese fries were nice and not too greasy!
<p6,e_3> The fries were perfect, cooked perfectly, not too hard, not too soft.
<p7,e_3> They have the best burgers, made with natural toppings , and best lettuce and tomatoes.
Lexicon
Postings
burger <p1,e_1> <p2,e_1> <p3,e_1> <p4,e_2> <p5,e_2> <p7,e_3>
toppings <p3,e_1> <p7,e_3>
price <p1,e_1> <p2,e_1>
fries <p1,e_1> <p2,e_1> <p5,e_2> <p6,e_3>
natural <p7,e_3>
… …
burger <p1,e_1> <p2,e_1> <p3,e_1> <p4,e_2> <p5,e_2> <p7,e_3>
fries <p1,e_1> <p2,e_1> <p5,e_2> <p6,e_3>
burger, fries <p1,e_1> <p2,e_1> <p3,e_1> <p4,e_2> <p5,e_2> <p6,e_3> <p7,e_3>
Rank 1
(e_1) Shake Shack
<p1,e_1>
<p2,e_1>
<p3,e_1>
Rank 2
(e_2) Umami Burger
<p4,e_2>
<p5,e_2>
Rank 3
(e_3) Burger Joint
<p6,e_3>
<p7,e_3>

Figure 1: Smith overview.

The retrieved passages are then grouped by their referencing entities. Finally, the scores of the passages are aggregated to compute the scores of the corresponding entities.

2. SMITH OVERVIEW

Figure 1 illustrates an overview of a search in Smith and highlights the key steps in Smith indexing and query processing. Let us start by assuming that we have three burger joints: Shake Shack (e1), Umami Burger (e2), and Burger Joint (e3) (Figure 1(a)). Smith retrieves reviews and comments about the restaurants from relevant sources such as Yelp and TripAdvisor, and partitions each review document into an array of short passages (1-2 sentences). Each passage is then parsed using a dependency parser. Using the resulting parse tree and dependency structure, information about negation and semantic relations between aspects (e.g., burger, price, parking) and sentiment words (e.g., best, delicious, decent) is extracted.

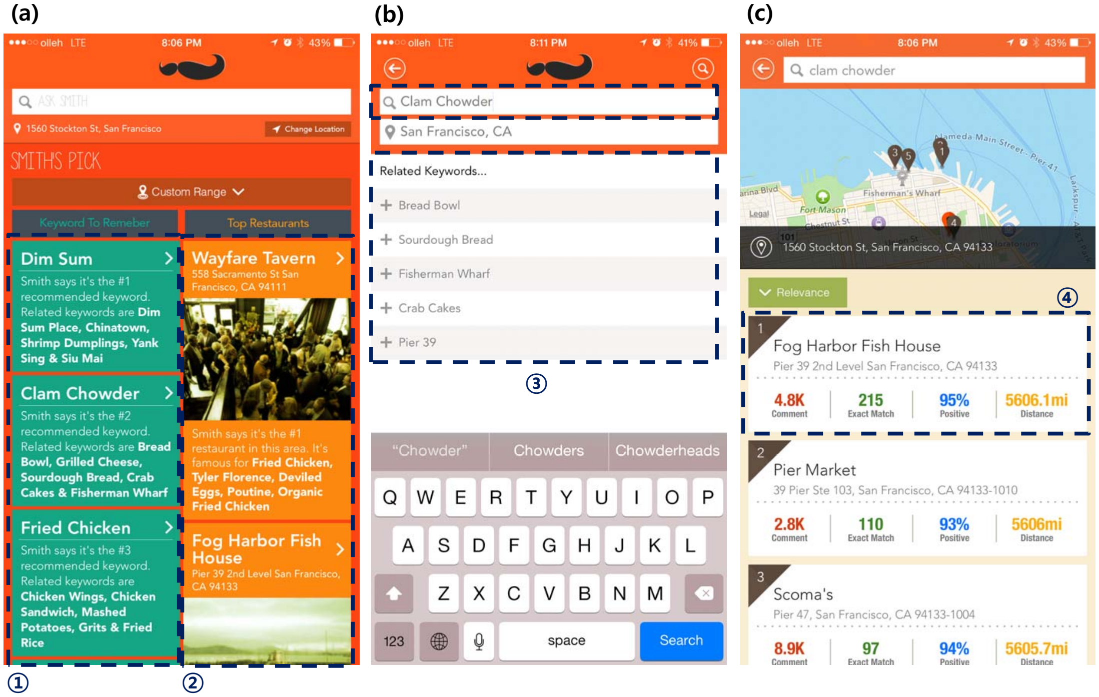
For example, in the second passage (p2), the user expressed positive opinions about Shake Shack's burgers and fries using sentiment words such as "good" and "delicious." Smith indexes these types of information, i.e., aspect-sentiment relations, along with other information including negation and entities described in the passage (Figure 1(b)). However, for brevity, in Figure 1, we omit the details and show only the passage ID and the entity ID in a posting (e.g., <p1, e1> represents that passage p1 describes restaurant e1). For indexing, Smith uses the conventional inverted index scheme where a term in the corpus is linked to a posting list that consists of the postings of the passages that contain the term. Figure 1(b) shows an example of the index. From the example, we know that the two passages p3 and p7, both of which contain the term "toppings," describe restaurants e1 and e3, respectively.

Given the user query, "best burgers and fries," Smith retrieves passages that match the query keywords (Figure 1(c)). Figure 1(d) illustrates two posting lists that match "burger" and "fries," respectively. For simplicity, let us assume that we process the query using only the two query terms. Matching passages are scored and grouped by entities that they describe. Finally, the entities are ranked according to the aggregate scores of the passages that describe them (Figure 1(e)). The passage scoring considers multiple factors including the amount of overlap with the query, negation, sentiment orientation and strength, authority of reviewer and site, review quality, and recency. For more details on the scoring metric, please refer to [1].

3. DEMONSTRATION

Figures 2 and 3 show the examples of the Smith result pages. When a user opens the application, Smith shows the selected keywords and top-rated restaurants near the user's current location. The keywords and restaurants are selected based on the sentiment scores that are associated with them. We use North Beach, San Francisco (US) for illustration.

Figure 2(a1) shows the top keywords for North Beach. Keywords are basically phrases that are frequently and positively mentioned within a selected area. In this example, North Beach is famous for "Dim Sum," "Clam Chowder," and "Fried Chicken." Related keywords are phrases that frequently co-occur with the top keywords. The keywords "Dim Sum Place," "Chinatown," "Shrimp Dumpling," "Yank Sing," and "Sui Mai" are related to the original keyword "Dim Sum." Figure 2(a2) shows the top-rated restaurants within the area. A top-rated restaurant is selected mainly by the number of times it is positively referenced in the review corpus. Smith recommends Wayfare Tavern and Fog Harbor Fish House if you are in North Beach.



Smith says it's the #1 recommended keyword. Related keywords are **Dim Sum Place, Chinatown, Shrimp Dumplings, Yank Sing & Siu Mai**

Smith says it's the #2 recommended keyword. Related keywords are **Bread Bowl, Grilled Cheese, Sourdough Bread, Crab Cakes & Fisherman Wharf**

Smith says it's the #3 recommended keyword. Related keywords are **Chicken Wings, Chicken Sandwich, Mashed Potatoes, Grits & Fried Rice**

Smith says it's the #1 restaurant in this area. It's famous for **Fried Chicken, Tyler Florence, Deviled Eggs, Poutine, Organic Fried Chicken**

1 Fog Harbor Fish House
Pier 39 2nd Level San Francisco, CA 94133
4.8K Comment | 215 Exact Match | 95% Positive | 5606.1mi Distance

2 Pier Market
39 Pier Ste 103, San Francisco, CA 94133-1010
2.8K Comment | 110 Exact Match | 93% Positive | 5606mi Distance

3 Scoma's
Pier 47, San Francisco, CA 94133-1004
8.9K Comment | 97 Exact Match | 94% Positive | 5605.7mi Distance

Figure 2: Smith Searches Restaurants: Home and List pages.

Figure 2(b) shows Smith's search page. Smith provides two sets of keywords: recommended keywords and related keywords. Before users start typing their query, Smith shows a list of recommended keywords that are computed in the same way as the top keywords on the front page (Figure 2(a1)). Once users input a query, Smith provides a set of keywords that are related to the user's query. In Figure 2(b3), the user searches "Clam Chowder" and Smith provides several new keywords related to clam chowder: "Bread Bowl," "Sourdough Bread," "Fisherman's Wharf," "Crab Cakes," and "Pier 39." This feature helps users navigate through the keyword space and helps quickly and easily familiarize users with the neighborhood.

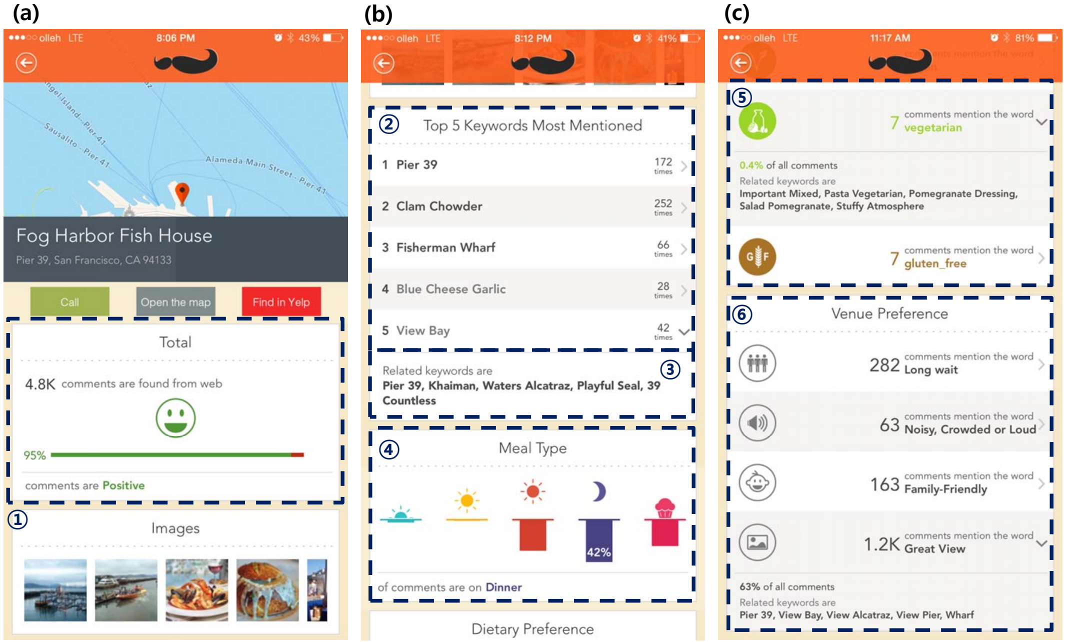
Figure 2(c4) shows the ranked list of restaurants for the query "clam chowder." Fog Harbor Fish House is returned as the most popular restaurant for the query. There are more than 4,800 comments on the restaurant and 215 positive opinions on its clam chowder. The distribution of positive and negative opinions in the 4,800 comments for Fog Harbor Fish House shows that 95% of people agree that Fog Harbor Fish House is a good restaurant while about 5% of people disagree. The distance is from the user's current location to the restaurant. Please note that we took the screenshot in Seoul and hence the distance shown in Figure 2(c4) is the distance from Seoul to North Beach, San Francisco.

When users click on a particular restaurant in the list, Smith provides the quantitative summary of opinions about the restaurant. Figure 3 shows the "detail" page of Fog Harbor Fish House. Figure 3(a1) presents the distribution of positive comments and Figure 3(b2) shows the popular restaurant keywords. We use TF-IDF weighting to rank the keywords. From these keywords, users can immediately see the restaurant's key features such as its prime location at Pier 39 in Fisherman's Wharf, its famous clam chowder and blue cheese garlic, and its view of the bay. When users click on a particular keyword, a set of related keywords is shown. For example, when users click on "bay view" to find more information about it, they will discover that they can also see Alcatraz and playful seals from the restaurant.

Figure 3(b4) shows the distribution of comments about "breakfast," "brunch," "lunch," "dinner," and "dessert." From this result, users know that this restaurant primarily serves lunch and dinner, and offers pretty good desserts. Figure 3(c5-6) shows seven aspects users are mostly interested in: three, related to dietary aspects, are "vegan," "vegetarian," and "gluten-free," and the rest, related to venue aspects, are "long wait," "noise level," "family-friendly," and "view." From this result, we know that Fog Harbor Fish House is unpopular with vegans and vegetarians and does not offer gluten-free meals. We also know that users should wait for a long time to be seated and that the restaurant has a great view and a family-friendly ambience.

4. SYSTEM SETTING

The current prototype of Smith was built using the review corpus crawled from Yelp and TripAdvisor. Smith indexes all reviews of California (CA), which are available at Yelp and TripAdvisor as of October 2014. A total of 67,144 restaurants are indexed. The number of restaurant review passages is 21,400,340, which were extracted from 6,346,684 original review posts. The size of index for Smith is 17 GB. Smith is running on a 4-node cluster on Amazon Web Service, where each node consists of a 2.56 GHz Intel Xeon E5-2670 v2 processor, 16 GB of memory, and 64 GB of SSD.



(a)

Fog Harbor Fish House
Pier 39, San Francisco, CA 94133

Call | Open the map | Find in Yelp

Total
4.8K comments are found from web
95%
comments are Positive

Images

(b)

Top 5 Keywords Most Mentioned

1	Pier 39	172 times
2	Clam Chowder	252 times
3	Fisherman Wharf	66 times
4	Blue Cheese Garlic	28 times
5	View Bay	42 times

Related keywords are
Pier 39, Khaiman, Waters Alcatraz, Playful Seal, 39 Countless

Meal Type
42%
of comments are on Dinner

Dietary Preference

(c)

7 comments mention the word vegetarian
0.4% of all comments
Related keywords are
Important Mixed, Pasta Vegetarian, Pomegranate Dressing, Salad Pomegranate, Stuffy Atmosphere

7 comments mention the word gluten_free

Venue Preference

282 comments mention the word Long wait
63 comments mention the word Noisy, Crowded or Loud
163 comments mention the word Family-Friendly
1.2K comments mention the word Great View
63% of all comments
Related keywords are
Pier 39, View Bay, View Alcatraz, View Pier, Wharf

Figure 3: Smith Searches Restaurants: Detail page.

Although Smith is ready to serve all requests, our search interface (app) is available for only iOS at the time of writing. We plan to announce the Android version of the app shortly. Smith will eventually grow to include blogs and other forms of social media, and cover all over the U.S. soon.

5. RELATED WORK

There exists a large body of research on entity linking, retrieval and semantic search [4, 3]. Among them, Ganesan and Zhai's Opinion Expansion (OE) and Query Aspect Modeling (QAM) [2] approaches are most relevant to ours. Both approaches concatenate all the reviews–on an entity–in a single document, and index the document using a standard text retrieval system. At query time, the OE expands a user's query using a predefined set of synonyms of opinion words, and processes the expanded query as usual. The QAM, an additional improvement, splits a query based on the aspects, processes each subquery separately, and aggregates the scores from the subqueries for a final computation of rankings. Finally, the ranks of the returned documents represent the ranks of the corresponding entities. The OE expands an opinion word such as "good" or "nice" in the query to a predefined set of 35 positive sentiment words, and expands an intensifier such as "very" to a collection of 21 similar adverbs. It appears that the expanded words completely dominate the aspect (or any context) words in the matching process, which produces a generic ranked result that does not change much with different queries. More importantly, systems based on a traditional search engine (e.g., OE and QAM) are incapable of producing textual and statistical summaries at query time. In our previous work, we showed that CONSENTO outperformed OE and QAM by substantial margins [1].

Commercial services such as Yelp and TripAdvisor use their own proprietary ranking methods. Although it would be interesting to compare Smith's ranking performance with theirs, it would require a large user study which we leave for future work.

6. CONCLUSION

In this work, we introduced Smith, a novel opinion-based restaurant search engine. We applied it to reviews crawled from popular web sites, which demonstrates its efficacy. In future work, we plan to expand our search engine to more domains including, for example, products, events, organizations, and social issues.

7. ACKNOWLEDGMENTS

This work was supported by the NRF of Korea grant (No. 2014R1A2A1A10051238), Commercializations Promotion Agency for R&D Outcomes grant funded by the Ministry of Science, ICT and Future Planning, and Business for Cooperative R&D between Industry, Academy, and Research Institute grant funded by Korea SMBA (No. C0208085).

8. ADDITIONAL AUTHORS

Youngjae Choi (Tgrape, Ltd., email: `yj.choi@tgrape.com`)

9. REFERENCES

[1] J. Choi, D. Kim, S. Kim, J. Lee, S. Lim, S. Lee, and J. Kang. Consento: A new framework for opinion based entity search and summarization. In *CIKM*, 2012.

[2] K. Ganesan and C. Zhai. Opinion-based entity ranking. *Information Retrieval*, 15(2), 2012.

[3] E. Meij, K. Balog, and D. Odijk. Entity linking and retrieval for semantic search. In *WSDM*, 2014.

[4] W. Shen, J. Wang, and J. Han. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE TKDE*, 27(2), 2014.