Finding the Right Social Media Site for Questions

Contents

1. Introduction

Introduction

- Topic specialization through the nature of the social sites is an extremely challenging problem.
  - Users questions are always short since they are not clear of what exactly their questions mean.
  - New social media sites are constantly being created and existing social media sites are constantly changing to match new technology trends which opens up a great challenge, capturing the dynamic of an extremely large number or quickly evolving social media sites in time.
  - For a social media site, we can obtain its content through several sources and methods. However, this content maybe highly conflicted, e.g., the frequency estimation of a single words maybe the same, similar, or totally different. How to combine this highly conflicting content poses a serious challenge.

Introduction

- The paper provides following framework for topic specialization:
  - Provides a framework for topic specialization by ranking sites for a given question by matching the content of the question and the site, rather than the characteristics of users and their ties.
  - We propose a novel method to understand a user's short question. Based on Wikipedia, after extracting keywords of a given question, we can expand each keyword.
  - We propose a novel method to explore the nature of social sites. Based on the discovered content preference, we can explain the highly variable Q&A behavior among social media sites.

2.
Topic Specialisation through Semantic Knowledge

Topic Specialisation through Semantic Knowledge

- The general **problem** we address can be formulated as a ranking task :
  - **Input**: Given a specific question, Q, and a set S, where each $s_i \in S$ is a social media site.
  - **Output**: A ranked version S of the social media site set S where each $s_i \in S$ is ranked according to its likelihood to give a response to question Q.
- The basic outline of the approach is as follows:
  - First we need to expand each question using Wikipedia.
  - Capture and Understand the content of social media sites using search engines.
  - Rank sites based on matching between expanded question and site contents.

Question Modelling

- ➢ Users usually have a rough idea of what exactly their questions are. So we need to expand questions.
- ➢ This is done as follows:
    - **Step I**: Extract keywords. For each question, extract its top ranked words as keywords. In this work, we select all nouns as keywords.
    - **Step II**: Expand keywords. For each keyword, we expand using its Wikipedia article obtained via API Interface.
    - **Step III**: Vector indexing. Index all returned Wikipedia articles as a question profile vector.
- ➢ As a whole, given a question q, we extract its keywords and query them on Wikipedia, then index these keywords as the word frequency vector $W(q) = \{p(w_1|Wiki) \cdots, p(w_m|Wiki)\}$.
- ➢ A long Wikipedia-based profile rather than an short question is submitted to Q&A system, which can represent the user's intent more effectively. As a result, a better modeling of users' question can be achieved.

Modeling a Social Media Site

- We capture and understand the content of social sites through the lens of search engines, which crawl the most popular, or representative content of social media sites
- We model the social Media site as follows:
    - **Step I**: Crawl content. For a candidate social site, we obtain its n most popular pages by searching with the empty string and restricting the domain to the subject site.
    - **Step II**: Vector indexing. Index all returned web pages as the social site profile vector.
- As a whole given a social media site $s\ i \in S$, $T(s_i, g_j, n) = \{p(w_1 | s_i, g_j, n), \cdots, p(w_m | s_i, g_j, n)\}$ is the k-dimensional word frequency vector of the top-n indexed pages return by search engine g_j within site s_i's domain.
- Here we are trying to model the site in terms of its most representative content which overcomes the current representations of social media sites.

Ranking Sites by Combined Searching

- The cosine distance is used to measure the similarity.
- However, the similarity measure D(si, q) implies that the similarity estimation is varied for different search engines, gj , and pages of the index n.
- Experiments suggests that all n>=5,are reasonable and thus the results are insensitive to n.
- We define the optimal estimation over all different search engines as:

$$p(w_j^o) = S(p(w_j^1), p(w_j^2), \ldots) + C(p(w_j^1), p(w_j^2), \ldots) \quad (1)$$

where S(.) is the sharing measure function to estimate the common shared belief between multiple sources and C(.) is the conflict function to measure and allocate the conflicting (nonshared) belief.

Overall Picture

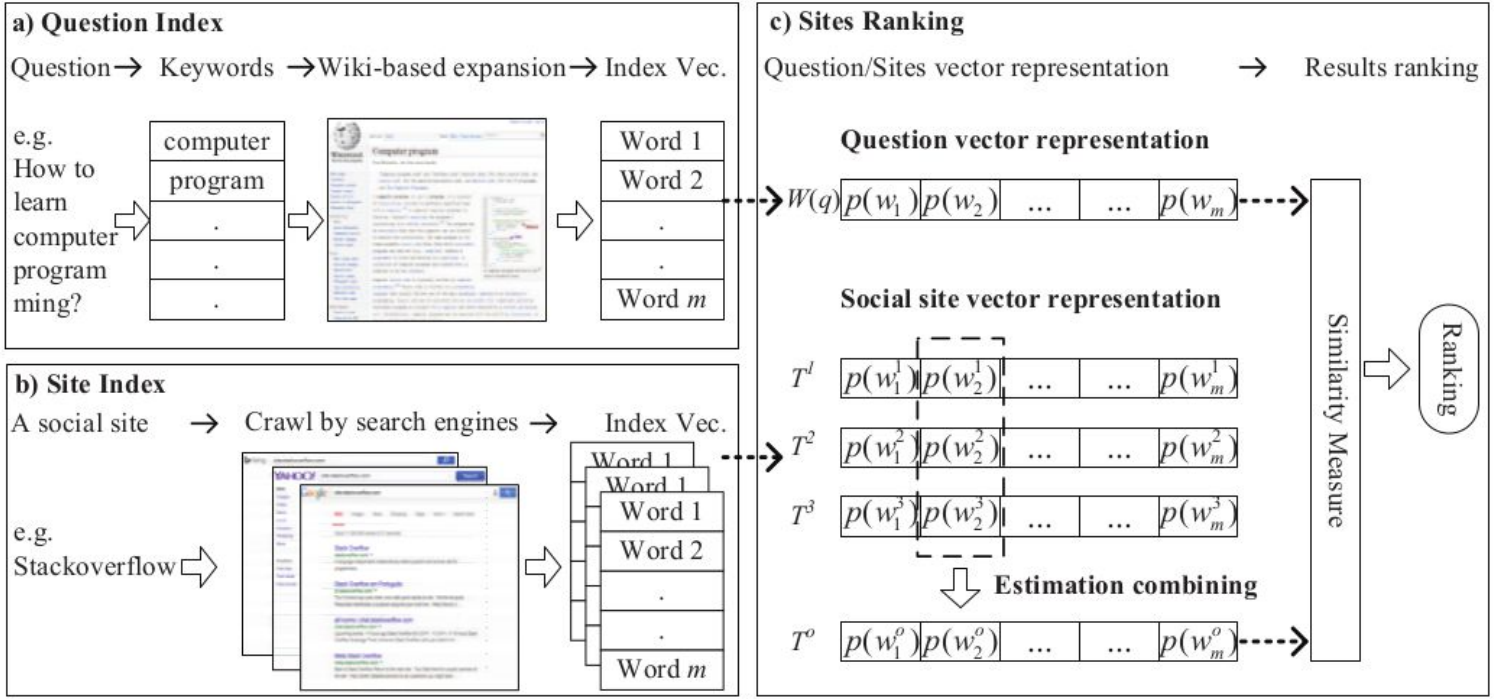

Fig. 1. Diagram of ranking sites for questions. (a) Modeling a users' question, (b) modeling a social media site, and (c) ranking sites by combined searching.

- With different sharing measures and conflict allocation strategies, we proposed several evidence combination rules:

(a) Max evidence combination

$$p(w_j^o) = \max_i(p(w_j^i)) \tag{2}$$

(b) Min evidence combination

$$p(w_j^o) = \min_i(p(w_j^i)) \tag{3}$$

(c) Mean evidence combination

$$p(w_j^o) = \frac{1}{M}\sum_i(p(w_j^i)) \tag{4}$$

(d) Dempster-Shafer evidence combination (DS) [6]

$$p(w_j^o) = p(w_j^1) \oplus p(w_j^2) \oplus \cdots \oplus p(w_j^N) \quad (5)$$
$$= \frac{1}{1-K} \sum_{\cap_i w_j^i = w_j} \prod_i p(w_j^i)$$

where $K = \sum_{\cap_i w_j^i = \emptyset} \prod_k p(w_j^i)$.

(e) Yager evidence combination (Yager) [7]

$$p(w_j^o) = \sum_{\cap_i w_j^i = w_j} \prod_i p(w_j^i) \quad (6)$$

(f) Conflict combination (CA):

$$p(w_j^o) = \sum_{\cap_i w_j^i = w_j} \prod_i p(w_j^i) \quad (7)$$
$$+ q(w_j)(1 - \sum_j \sum_{\cap_i w_j^i = w_j} \prod_k p(w_j^i))$$

where $q(w_j) = \frac{\sum_i p(w_j^i)}{\sum_i \sum_j p(w_j^i)} = \frac{\sum_i p(w_j^i)}{M}$.

- The above equations considers only sharing function S(.) and simply ignores the conflict function C(.) which usually results in wrong results.
- we try to allocate the conflict probability proportionally. We let the sharing measure function as:

$$S = \sum_{\cap_i w_j^i = w_j} \prod_i p(w_j^i).$$

- Thus the conflicting probability should be:

$$1 - \sum_j \sum_{\cap_i w_j^i = w_j} \prod_k p(w_j^i).$$

Experiment Setup

A)**Data:**

- Two data sets were used for experimental evaluation in this work:

1)Selected Questions:10 questions were selected,which were the top 10 most asked topics on Internet.

2)Factoid Q&A Corpus:We used the factoid Q&A Corpus [11], which contain 1,714 manually-generated factoid questions and their corepondіng answers collected by Carnegie Mellon University and the University of Pittsburgh between 2008 and 2010.

Experiment setup:

B) i)**Candidate Sites:** We select 17 well known social media sites from the top 200 sites listed on Alexa.com as the candidate site set S. Also, we manually add 8 well known professional social media sites to the candidate site set S for some specific domains, such as Linkedin for job hunting, Match.com for dating, etc. In total there are 25 sites as candidate sites.

ii)**Ground Truth:**

- For a specific question, we crawl the Wikipedia articles for its profile per keyword.
- For these candidate sites, we also crawl their Wikipedia articles as profiles, then rank the sites by the cosine similarity.
- For a specific question, we use the article which its answer was extracted from as its profile.
- For candidate sites, we send them to search engine (e.g. Google) and crawl the top-5 returned pages in the site’s domain as profiles, then rank the sites by the cosine similarity.
- Since the answer articles are manually selected for the specific question, these ranking results S∗ were treated as the ground truth.

Evaluation Metric: For the ranked version S' of the candidate set S, we evaluate the performance by comparing S' and S^* using the top-n intersection rate, namely the fraction of the common elements in top n ranking results: $\frac{S'(n) \bigcap S^*(n)}{n}$, where the $S'(n)$ and $S^*(n)$ denote the top n ranking results of S' and S^*. Supposing that the top 10 ranked sites in S^* are correct answers of each question, we also can evaluate the top-n accuracy rate and average precision (AP) [12], which are widely used in IR and keyword evaluation.

THANK YOU