Finding Right Social Media for Questions

R Vamshi Teja
N L Rajesh
V Rajesh
K Vivek

IIT Hyderabad

April 17, 2018

Overview

1. Motivation
2. Previous approaches
3. Proposed Methology
   - Question Modelling
   - Site Modelling
   - Ranking Sites by Combined Searching
4. Experiments and Results
   - Dataset
   - Implementation Details
   - Evaluation Metric
   - Results
5. Future Scope of Improvement
6. References

Motivation

- Social media has become a part of our daily life. Most use it, some are developing a dependence on it, and a few make their living on it.
- Unfortunately, our questions donâĂŹt always receive satisfactory responses, which occurs partly because the responses are highly variable among social media sites, i.e., the quality of responses is quite different when asked on different social media sites.
- So there is a need to assist users in finding the right domain to get their questions answered.

Proposed Methodology

- The problem we are addressing can be formulated as a ranking task :
  - Given a specific question, Q, and a set S, where each $s_i \in$ S is a social media site.
  - 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.
- To expand questions following framework is proposed:
  - Extract keywords. For each question, extract its top ranked words as keywords. In this work, we select all nouns as keywords.
  - Expand keywords. For each keyword, we expand using its Wikipedia article obtained via API Interface.
  - 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) = $p(w_1|\text{Wiki}), \ldots, p(w_m|\text{Wiki})$.
- A long Wikipedia-based profile rather than an short question is submitted to Q&A system, which can represent the userâĂŹs intent more effectively. As a result, a better modeling of usersâĂŹ question can be achieved.

- Let us consider an example:
- 'How to learn computer programming?'
  - Nouns extracted from the above question are:
  - For each noun we will expand a Wikipedia article now.
  - Now top M words extracted from the articles and indexed for the question:

Site Modelling

- 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:
    - 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.
    - Vector indexing. Index all returned web pages as the social site profile vector.

- As a whole given a social media site s_i E S, T (s_i , g_j , n) = p($w_1|s_i, g_i, n$), ... , 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 siâĂŹ 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 similarity measure D(s_i, q) implies that the similarity estimation is varied for different search engines, g_j,and pages of the index n.
- Our experiments suggest that every reasonable n value (greater than 5) can work,and results are, insensitive to this choice.
- Hence,T(s_i,g_j,n) can be simplified as T^j = p(w_1^j), . . . ,p(w_k^j) , with j varied by using different search engines g_j.

- Hence optimal solution is T^o = p(w_1^o), . . . ,p(w_k^o) where p(w_j^o) is a combination of independent set of probability estimation of p(w_j^1),p(w_j^2), . . . ,p(w_j^M) observed by N search engines.
- However, since the p(w_j^i) is observed by different search engines, different search engines express different beliefs i.e., for a word, its frequency observation from different search engines may be different.

- We proposed a united framework to combine the evidence obtained by different search engines, and the optimal estimation p(w_j^o) can be denoted as:
  p(w_j^o) = S(p(w_j^1), p(w_j^2), ...) + C(p(w_j^1), p(w_j^2), ...)
  where S(.)-common shared belief between multiple sources
  C(.) is the conflicting (non-shared) belief.
- we proposed several evidence combination rules for shared belief:
  - (a) Max evidence combination:
    $$\mathrm{p}(w_j^o)=max_i(\mathrm{p}(w_j^i))$$
  - (b) Min evidence combination:
    $$\mathrm{p}(w_j^o)=min_i(\mathrm{p}(w_j^i))$$
  - (c) Mean evidence combination:
    $$\mathrm{p}(w_j^o) =(1/\mathrm{M})\sum_i(p(w_j^i))$$

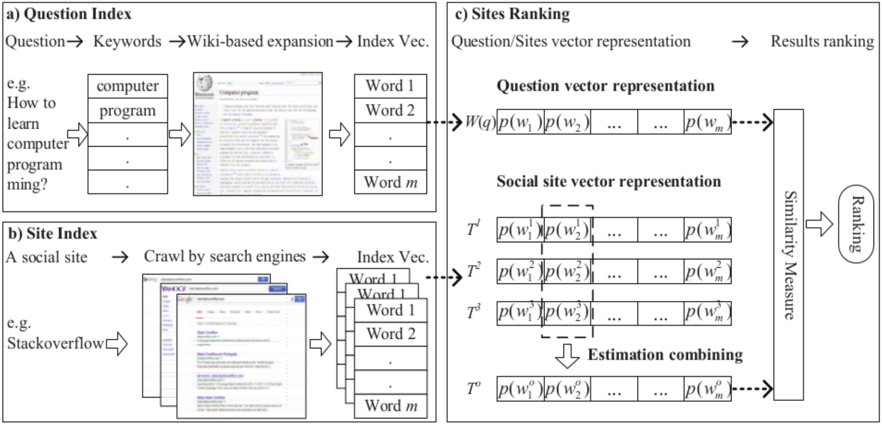
- Sites Ranking: Thus for question q, the site set S can be ranked by D(s_i,q)=(<T(s_i,g_j,n),W(q)>) / (||T(s_i,g_j,n)||x||W(q)||)
- Combination rules 1 through 3 only measure the shared belief between multiple sources, i.e., are kinds of sharing function S(.), and simply ignores all the conflicting belief or leverages it through a normalization factor, i.e., the conflicting allocation function C(.) = 0. This simplification produces wrong results in case of high conflict.
- If non-zero conflict function is considered,then we let shared function to be:

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

  The conflict function is given by:

  $$\text{C}=1\text{-}\sum_j\sum_{\cap_i\ w_j^i=w_j}\prod_i(p(w_j^i))$$

Overall Picture



a) Question Index
Question→ Keywords →Wiki-based expansion→ Index Vec.
e.g. How to learn computer program ming?
computer
program
Word 1
Word 2
Word m
b) Site Index
A social site → Crawl by search engines → Index Vec.
e.g. Stackoverflow
Word 1
Word 2
Word m
c) Sites Ranking
Question/Sites vector representation → Results ranking
Question vector representation
$W(q)$ $p(w_1)$ $p(w_2)$ ... ... $p(w_m)$
Social site vector representation
T^1 $p(w_1^1)$ $p(w_2^1)$ ... ... $p(w_m^1)$
T^2 $p(w_1^2)$ $p(w_2^2)$ ... ... $p(w_m^2)$
T^3 $p(w_1^3)$ $p(w_2^3)$ ... ... $p(w_m^3)$
Estimation combining
T^o $p(w_1^o)$ $p(w_2^o)$ ... ... $p(w_m^o)$
Similarity Measure
Ranking

Figure: Diagram of ranking sites for questions. (a) Modeling a usersâĂŹ question, (b) modeling a social media site, and (c) ranking sites by combined searching.

Dataset

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

- Selected Questions:10 questions were selected,which were the top 10 most asked topics on Internet.
- Factoid Q&A Corpus:We used the factoid Q&A Corpus, which contain 1,714 manually-generated factoid questions and their coreponding answers collected by Carnegie Mellon University and the University of Pittsburgh between 2008 and 2010.

Implementation Details

- For question modelling
  - we have used wikipedia api to expand the query
  - wee have used vocabulary size of m = 400000
- For site modelling
  - We have selected 26 sites as shown in figure from different categories. Distribution of sites category wise is also shown in figure

Evaluation Metrics

- We use top-n intersection rate as our evaluation metric.

$$TopnIntersection = \frac{|S(n) \cap S^*(n)|}{n}$$

where S(n) and $S^*(n)$ denote top n ranking results of S and S^*.

- We also used average precision as our evaluation metric.

Results

Scope of Improvement

- We can Incorporate *Deep Semantic Similarity Model(DSSM)* to improve ranking.
- Current framework is based on static links(site modelling). Dynamic modelling of sites can help.

References

📄 Zhen Yang, Isaac Jones , Xia Hu , Huan Liu. *Finding the Right Social Media Site for Questions.*