Finding High-Quality Content in Question-Answering Portal

Aparna Todwal[A], Prof. M. Wanjari[B]

[A] CSE Department, SRCOEM, Maharashtra, India, todwal.aparna@gmail.com
[B]Associate Professor, CSE Department, SRCOEM, Maharashtra, India,wanjarimr@rknec.com

Abstract

In these recent years, interest of people in social networking sites is flowing in a direction towards generating data. So the quality of user-generated content changes drastically from Very good to abuse and bad. As the presence of such content develops, the task of identifying high-quality content in sites based on user contributions in social media sites which becomes drastically crucial. Social media in general serves a rich range of information sources: in addition to the content itself, there is a broad range of non-content information present, such as association between items and explicit quality ratings from team-worker of the community. This paper finds methods for exploiting such community feedback to automatically identify high quality content. As a test case, the concern is toward Yahoo! Answers system, a large community question/answering portal. In particular, for the community question/answering domain, this paper shows that our system is able to find high-quality items from the rest.

Keywords: *Social media, Community Question Answering, User Interactions..*

I. INTRODUCTION

From the early 2000s, user-generated contents has become increasingly famous on www; so majority of users participate in content Creation, rather than just using it. Popular social media domain contains various blogs system, social bookmarking sites, photo and video sharing communities and social networking platform such as MySpace and Facebook, which offers a combination of all these functionalities and relationship among users. Community-driven question/answering portals are the some different form of user-generated content that is gaining a large audience attraction in recent years. These systems provide some other direction for gaining information on the web. An important differentiation between content generated by and traditional content that is particularly significant for knowledge-based media such as question/answering portals is the difference in the quality of the content quality. The important hurdle posed by content in social media sites is the fact that the distribution of quality has high variance: from very high-quality items to not related quality items, sometimes very bad content. This makes the program of filtering and ranking in such systems more complex than in other domains.

II. BASIC ELEMENTS OF QUESTION/ANSWERING PORTAL

In this paper we address the task of identifying high-quality content in community-driven question/answering sites. As a test case, we focus on Yahoo! Answers, a large portal that is particularly rich in the amount and types of content and social interaction available in it. We focus on the following research questions:

1. *What are the different elements of social media that can be utilized to facilitate automated discovery of good-quality content?* In sum to the content itself, there is a very broad range of non-content information present, from the link between items to explicit and implicit quality rating from members of the community. What is the utility of each source of information to the task of estimating quality?

2. *How are these different factors associated? Is content alone enough for identifying high-quality items?*

3. *Can community feedback averages judgments of specialists?*

A. Yahoo! Answers

Yahoo! Answers is a question/answering system where users ask the question and answer questions on any matter. A user can mark the answers good or bad accordingly of other users, mark good questions and answers. Thus, overall, each user has a three tasks: asker, answerer and evaluator.

The major element of the Yahoo! Answers system are basically questions. All the questions has its lifecycle. It starts in an “open" state where it gets the answers. Then at a point the question is known to be closed and can get no answers. At this point, a “best answer" is elected either by the asker or with the help of voting procedure from other users; once a good quality answer is selected, the question is resolved. Yahoo! Answers is a very famous service as a result, it hosts a very broad amount of questions and answers in a wide variety of areas, making it a specifically useful area for evaluating content quality in social media

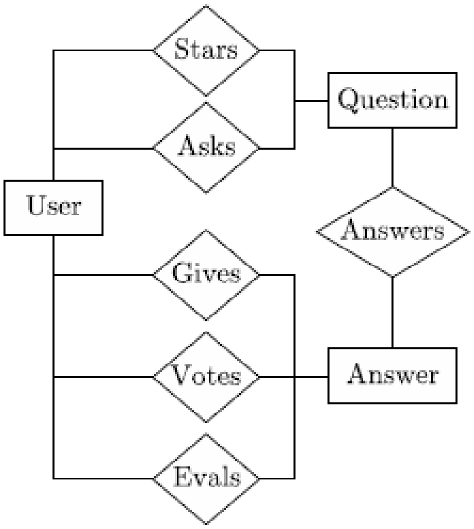


Fig:-1 diagram of relationship of users with answers of question.
(Adopted from [1])

Now, in a system where multiple users gives stars to the question and can also asks the question. Again user can also give their answer to the question, can mark the answers as good or bad and also can estimates the answer. The correlation between questions, users those who asks and those who answers questions, and answers can be extracted by a tripartite graph given in Figure 2, different links between the nodes gives an explicit link between various node types

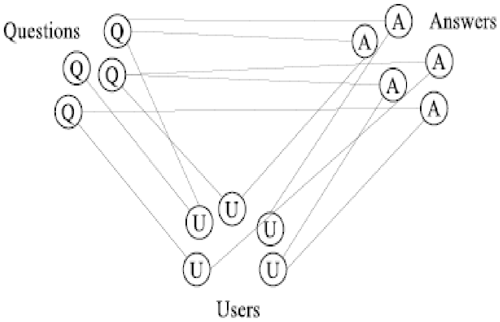


Fig:-2 Association of users-question-answers modeled as tri-partite graph. (Adopted from [1])

As shown in figure 2, a user is not supposed to answer his/her own questions, there are no triangles in the given diagram, so all cycles in the graph have their length at least 6.[1]

III. RELATED WORK

A. Link analysis in social media

Specifically, link-based ranking algorithms that were triumphant in finding the quality of web pages have been applied in this context. Link-based methods have been shown to be triumphant for many of tasks in social media [2].Two of the most proper and useful link-based ranking algorithms are Page Rank [4] and HITS [3].

Consider a graph $G = (V;E)$ with vertex set V corresponds to the users of a question/answer system and contains a directed edge $e = (u; v) \in E$ from a user $u \in V$ to a user $v \in V$ if user u has answered to at least single question of user v. Expertise Rank [8] corresponds to Page Rank over the transposed graph *G*□= (*V;E*□), that is, a score is cultivated from the person getting the answer to the person giving the answer. The repetition implies that if person u was capable to give an answer to person v, and person v was able to provide an answer to person w, then u should receive some added points given that he/she was able to provide an answer to a person with a certain level of expertise.

B. Propagating reputation

Guha et al. [5] study the problem of proliferating trust and distrust among Epinions users, who may gives positive (trust) and negative (distrust) ratings to each other. The authors study ways of averaging trust and distrust and observe that, while auditing trust as a transitive property makes sense, distrust cannot be taken as transitive.

Ziegler and Lausen [7] has also studied the models for proliferation of trust. They shows a anatomy of trust metrics and discuss ways of adding information about distrust into the rating scores.

C. Question/answering portals and forums.

The specific framework of question/answering communities we concentrate on in this paper has been the object of some study in recent years.

According to Su et al. [6], the quality of answers in question/answering portals is better on average, but the quality of particular answers changes considerably. Specifically, in a study of the answers to a set of questions in Yahoo! Answers, the authors got that the portion of correct answers to particular questions asked by the authors of the study, varied from 17% to 45%. The portion of questions in their sample with atleast one better answer was much greater, changing from 65% to 90%, meaning that a method for estimating high-quality answers can have a good effect in the user's satisfaction with the system.

D. Expert Finding.

Zhang et al. [8] evaluates data from an online forum, searching to identify users with high expertise. They investigate the user answers graph in which there is relation between users u and v if u answers a question by v, referring both Expertise Rank and HITS to find users with high expertise. Their output provides high relationship between link-based metrics and the answer quality. The authors also builds synthetic models that record some of the characteristics of the interactions among users in their dataset.

The HITS algorithm is work on the user-answer graph. Jurczyk and Agichtein [9] shows an application of the HITS algorithm [3] to a question/answering portal. The output estimates that HITS is a favorable approach, as the obtained authority score is good related to the number of votes that the items receive, than simply estimating the number of answers the answerer has given in the past. Dom et al. [10] studied the judgment of several link-based algorithms to rank people by expertise on a network of e-mail exchanges.

IV. PROPOSED ARCHITECTURE

In the literature survey, we came across many ways by which the content quality can be estimated in the question answering portal on the basis of feedbacks registered by different users.

But the problem involved in such a scheme is that the users feedback is mandatory in getting the ratings to the answer so that finding relevancy among the answers

Here is an attempt to obtain the most relevant answers among all the given answers without user's user feedback registration.

For this purpose, as we are dealing with the community driven question answering portal where for multiple question and answers can be given so we are using the concept of calculating the scores for each of the answers given for a question. After having scores for each of the answers and scores can be generated by applying some mechanism to it will be explained letter. Now it is easy to estimate the high quality content among all the answers by taking the high scorer answer on the top. There are some parameters on which the task will accomplish.

V. IMPLEMENTATION DETAILS

The above parameter was implemented with the help of Wordnet and OpenNLP toolkit , the GUI was built in Microsoft Visual Studio 2010 and the language used is c#.

The flow of project is described below.

1) User login Registration
2) Adding question
3) Standard answer generation
4) Score calculation by comparing answer to that of standard answer with the help of Wordnet
5) Users rating can also be taken into consideration for finding relevant contents on question answering portal.

6) Finally high quality answers (Relevant) gets the top most position according to their scores .

Before the implementation is started we required Natural Language Processing Tool kit (OpenNLP). The another package required is the Wordnet for the comparison purpose and score calculation.

The steps implemented are as follows:

A. User login Registration

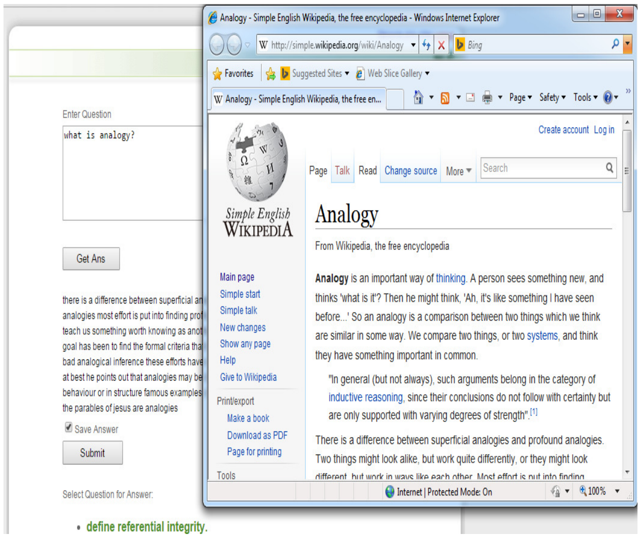
- As we are considering for finding relevant contents on question answering portal.
- So multiple users take part in rating the answers
- So it is required to register users.

B. Adding question

- After user login the question gets added

C. Standard answer generation

- As the question gets entered then the browser pops up and the information related to the keyword in question is gets extracted
- Then the standard answer gets generated.



Enter Question

what is analogy?

there is a difference between superficial an
analogies most effort is put into finding prof
teach us something worth knowing as anot
goal has been to find the formal criteria tha
bad analogical inference these efforts have
at best he points out that analogies may be
behaviour or in structure famous examples
the parables of jesus are analogies

☑ Save Answer

Select Question for Answer:

- define referential integrity.

Analogy

From Wikipedia, the free encyclopedia

Analogy is an important way of thinking. A person sees something new, and thinks 'what is it'? Then he might think, 'Ah, it's like something I have seen before...' So an analogy is a comparison between two things which we think are similar in some way. We compare two things, or two systems, and think they have something important in common.

> "In general (but not always), such arguments belong in the category of inductive reasoning, since their conclusions do not follow with certainty but are only supported with varying degrees of strength".[1]

There is a difference between superficial analogies and profound analogies. Two things might look alike, but work quite differently, or they might look different, but work in ways like each other. Most effort is put into finding

Fig: 3 Standard answer generations

D. Score calculation by comparing answer to that of standard answer with the help of Wordnet

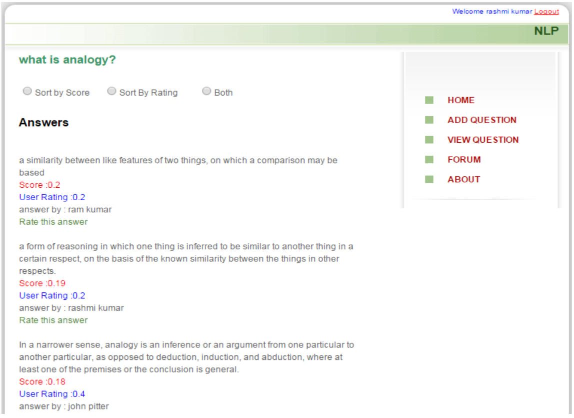
- Tokens form the standard answer is taken out
- Then the tokens from the answer is taken out.
- Comparison is done with the help of Wordnet database
- Score calculation is done with the help of wordnet

E. *Users rating can also be taken into consideration for finding relevant contents on question answering portal.*

- Users can register their rating by selecting a option among all the given options.
- User rating score gets evaluated

F. Finally high quality answers (Relevant) gets the top most position according to their scores.

- Once all the calculation done for the purpose of score done.
- The highest score answer gets on the top most position



what is analogy?

Answers

a similarity between like features of two things, on which a comparison may be based
Score :0.2
User Rating :0.2
answer by : ram kumar
Rate this answer

a form of reasoning in which one thing is inferred to be similar to another thing in a certain respect, on the basis of the known similarity between the things in other respects.
Score :0.19
User Rating :0.2
answer by : rashmi kumar
Rate this answer

In a narrower sense, analogy is an inference or an argument from one particular to another particular, as opposed to deduction, induction, and abduction, where at least one of the premises or the conclusion is general.
Score :0.18
User Rating :0.4
answer by : john pitter

Fig: 4 High score answer on top position

VI. CONCLUSION AND FUTURE WORK

Mobile agents move the data to the remote distributed databases, not the entire databases to the data. Therefore, the proposed system has huge bandwidth savings and can overcome network latency. Mobile Agents have received a diverse range of applications in information retrieval, network management, e-commerce, transportation systems, Distributed control systems, and manufacturing. From the literature survey it was noted that mobile agents have several advantages over conventional client server paradigm like reducing network traffic, supporting disconnected operation, overcoming network latency, and roaming ability in heterogeneous platforms, which is vital in building ubiquitous e-commerce systems. As compared to traditional client server systems MAs also provide fast and efficient interaction in an emerging e-commerce model.

For future we can implement a separate database for the mobile agent at the buyers end so that the searching time can be reduced to a certain amount.

ACKNOWLEDGMENT

The authors are grateful to the Head, Computer Science and Engineering Department and the Principal, SRCOEM (Shri Ramdeobaba College of Engineering and Management) for providing adequate facilities to conduct a research. Authors are thankful to the faculty members of Computer Science and Engineering Department for their support and cooperation during this work

.

REFERENCES

[1] Carlos Castillo, Debora Donato et al., “FINDING HIGH QUALITY CONTENTS IN SOCIAL MEDIA”Yahoo! Research Barcelona, Spain, WSDM'08, February 11.12, 2008.

[2] J. P. Scott. Social Network Analysis: A Handbook. SAGE Publications, January 2000.

[3] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5):604{632, 1999.

[4] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: bringing order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998.

[5] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. In WWW '04: Proceedings of the 13th international conference on World Wide Web, pages 403{412, New York, NY, USA, 2004. ACM Press.

[6] Q. Su, D. Pavlov, J.-H. Chow, and W. C. Baker. Internet-scale collection of human-reviewed data. In WWW '07: Proceedings of the 16th international conference on World Wide Web, pages 231{240, New York, NY, USA, 2007. ACM Press.

[7] C.-N. Ziegler and G. Lausen. Propagation models for trust and distrust in social networks. Information Systems Frontiers, 7(4-5):337{358, December 2005.

[8] J. Zhang, M. S. Ackerman, and L. Adamic. Expertise networks in online communities: structure and algorithms. In WWW '07: Proceedings of the 16th international conference on World Wide Web, pages 221-230, New York, NY, USA, 2007. ACM Press.

[9] P. Jurczyk and E. Agichtein. HITS on question answer portals: an exploration of link analysis for author ranking. In SIGIR (posters). ACM, 2007.

[10] B. Dom, I. Eiron, A. Cozzi, and Y. Zhang. Graph-based ranking algorithms for e-mail expertise analysis. In Proceedings of Workshop on Data Mining and Knowledge Discovery, pages 42{48, San Diego, CA, USA, 2003. ACM Press.

[11] T. Joachims, L. A. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting click through data as implicit feedback. In SIGIR, pages 154{161, 2005.

[12] K. Ali and M. Scarr. Robust methodologies for modeling web click distributions. In WWW, pages 511-520, 2007.

[13] C. Anderson. The Long Tail: Why the Future of Business Is Selling Less of More. Hyperion, July 2006.

[14] Y. Attali and J. Burstein. Automated essay scoring with e-rater v.2. Journal of Technology, Learning, and Assessment, 4(3), February 2006.

[15] A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra. A maximum entropy approach to natural language processing. Computational Linguistics, 22(1):39-71, 1996.

[16] Aparna Todwal and Prof. M. Wanjari “Finding High Quality Contents in Social Media”, In International Journal of Engineering Trends and Technology, Volume 21, number 6 –March 2015