Advanced Scalable Algorithm for Community Question Answering Using Post Voting Prediction

[1]Mr.G.Prem Paul , [2]Sricharukesh.E.G , [3]Vignesh Kumar.S ,[4]Gokul Kannan.R,
[1]Assistant Professor ,Department of Computer Science and Engineering
K.L.N. College of Engineering ,Madurai,India.
[234]Department of Computer Science and Engineering ,K.L.N. College of Engineering
Madurai,India.

ABSTRACT - *Community Question Answering (CQA) sites, such as Stack Overflow and Yahoo! Answers, have become very popular in recent years. These sites contain rich crowdsourcing knowledge contributed by the site users in the form of questions and answers, and these questions and answers can satisfy the information needs of more users. In this article, we aim at predicting the voting scores of questions/answers shortly after they are posted in the CQA sites. To accomplish this task, we identify three key aspects that matter with the voting of a post, i.e., the non-linear relationships between features and output, the question and answer coupling, and the dynamic fashion of data arrivals. A family of algorithms are proposed to model the above three key aspects. Some approximations and extensions are also proposed to scale up the computation. We analyze the proposed algorithms in terms of optimality, correctness, and complexity. Extensive experimental evaluations conducted on two real data sets demonstrate the effectiveness and efficiency of our algorithms.*

INTRODUCTION :

It is a process of inspecting, cleansing, transforming,and modelling dat a with the goal of discovering useful information, suggesting conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, in different business, science, and social science domains.

Data mining is a particular data analysis technique that focuses on modelling and knowledge discovery for predictive rather than purely descriptive purposes, while business intelligence covers data analysis that relies heavily on aggregation, focusing on business information. In statistical applications, data analysis can be divided into descriptive statistics, exploratory data analysis (EDA), and confirmatory data analysis (CDA). EDA focuses on discovering new features in the data and CDA on confirming or falsifying existing hypotheses. Predictive analytics focuses on application of statistical models for predictive forecasting or classification, while text analytics applies statistical, linguistic, and structural techniques to extract and classify information from textual sources, a species of unstructured data. All are varieties of data analysis. Data Science Process Flowchart From "Doing Data Science", Cathy O'Neil And Rachel Schutt, 2013Analysis Refers To Breaking A Whole Into Its Separate Components For Individual Examination. Data Analysis Is A Processors Obtaining Raw Data And Converting It Into Information Useful For Decision-Making By Users. Data Is Collected And Analyzed To Answer Questions, Test Hypotheses Or Disprove Theories. Statistician John Tukey Defined Data Analysis In 1961 As: "Procedures For Analyzing Data, Techniques For Interpreting The Results Of Such Procedures, Ways Of Planning The Gathering Of Data To Make Its Analysis Easier, More Precise Or More Accurate, And All The Machinery And Results Of(Mathematical) Statistics Which Apply To Analyzing Data.

II.SCOPE OF THE PROJECT:

Community Question Answering (CQA) sites have become valuable repositories that host a massive volume of human knowledge. In addition to providing answers to the questioner, CQA sites now serve as knowledge bases for the searching and browsing conducted by a much larger audience. For example, in software forum called Stack Overflow, programmers can post their programming questions on the forum, and others can propose their answers for these questions. Such questions as well as their associated answers could be valuable and reusable for many other programmers who encounter similar problems. In fact, millions of programmers now use such forums to search for solutions for their programming problems. To maximize the utility of CQA sites, a key task is to characterize and predict the intrinsic value (e.g., quality, impact, etc) of the question/answer posts. This is an essential task for both information producers and consumers. From the perspective of information producers (e.g., who ask or answer questions), it would be helpful to identify the high value questions in the early stage so that these questions can be recommended to experts for them to answer. From the Perspective of information consumers (e.g., who search or browse questions and answers), it would be helpful to highlight high-value questions/answers (e.g., by displaying them more prominently on the site or allowing the search engine to be aware of their value) so that users can easily discover them. Most of the existing CQA sites allow the site users to vote (e.g., up vote and down vote in Stack Overflow) for a question or an answer. The outcome of such voting, e.g., the difference between the number of the up votes and down votes that a question/answer receives from the site users (referred to as 'voting score'), provides a good indicator of the intrinsic value of a question/answer. To some extent, the voting score of a question/answer resembles the number of the citations that a research paper receives in the scientific publication domain. It reflects the net number of users who have a positive attitude toward the paper. In the past, the voting score has been studied in several interesting scenarios (e.g., information quality, user satisfaction, etc; see related work section for details). In this paper, we aim to study the relationship between the voting scores of questions and those of answers. We conjecture that there exists *correlation* between the voting score of a question and that of its associated answer. Intuitively, an interesting question might obtain more attention from potential answerers and thus has a better chance to receive high-score answers. On the other hand, it might be very difficult for a low-score question to attract a high-score answer due to, e.g., its poor expression in language, or lack of interestingness in topic. Starting from this conjecture, we study two real CQA sites, i.e., Stack Overflow1 (*SO*), and Mathematics Stack Exchange2 (*Math*). Our key finding is that the voting score of an answer is indeed strongly positively correlated with that of its question. Such correlation structure consistently exists on both sites. Armed with this observation, we propose a family of algorithms (*CoPs*) to *jointly* predict the voting scores of questions and answers. In particular, we aim at identifying the potentially high-score posts soon after they are posted in the CQA sites.

III.SYSTEM SPECIFICATIONS:

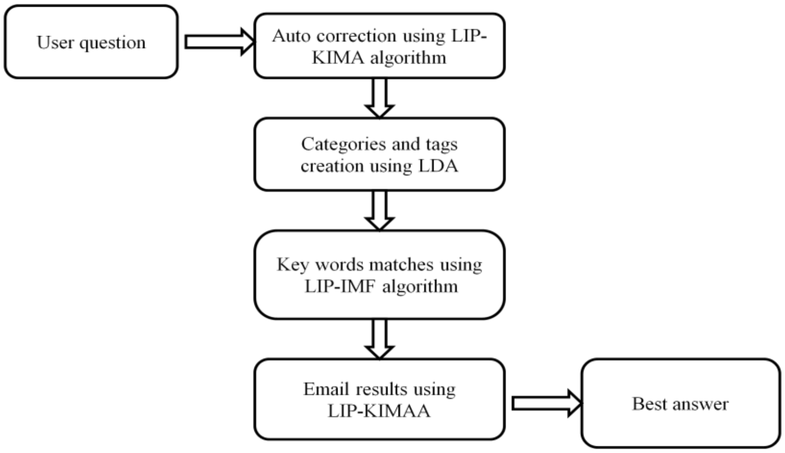
Hardware System Configuration:-

- Processor - INTEL Core i3-4160
- RAM – 8 GB
- Hard Disk - 500 GB

Software System Configuration:-

- Operating System: Windows 7
- Application Server : WAMP SERVER
- Front End: PHP
- Database: MySQL 5.0

IV.BLOCK DIAGRAM:


User question
Auto correction using LIP-KIMA algorithm
Categories and tags creation using LDA
Key words matches using LIP-IMF algorithm
Email results using LIP-KIMAA
Best answer

V. Existing System:

We focus on the voting score prediction of questions/answers shortly after they are posted in the CQA sites. Such a task is essential for the prosperity and sustainability of the CQA ecosystem, and it may benefit all types of users, including the information producers and consumers. For example, detecting potentially high score answers can benefit the questioners as well as the people who have similar questions; it would also be helpful to identify high-score questions in the early stage and route them to expert answerers.

VI.Proposed System:

We have proposed a family of algorithms to comprehensively and efficiently predict the voting scores of questions/answers in CQA sites. In particular, some of the proposed algorithms (LIP-KIM, LIP-KIMA, and LIP-KIMAA) can capture three key aspects (non-linearity, coupling,

and dynamics) that matter with the voting score of a post, while others can handle the special cases when only a fraction of the three aspects are prominent. In terms of computation efficiency, some algorithms (LIP-IM, LIP-IMF, LIP-KIA, LIP-KIMAA, and LIP-KIMAA) enjoy linear, sub-linear, or even constant scalability. The proposed algorithms are also able to fade the effects of old examples (LIP-IMF), and select a subset of features/examples (LIPMS and LIPKMS). We analyze our algorithms in terms of optimality, correctness, and complexity, and reveal the intrinsic relationships among different algorithms. We conduct extensive experimental evaluations on two real data sets to demonstrate the effectiveness and efficiency of our approaches.

VII.MODULES:

A) Points-based reputation management & auto correction on question using KIMA algorithm:

AUTOCORRECTION:

Text replacement, replace-as-you-type or autocorrect is an automatic data validation functioncommonly found in word processors and text editing interfaces for smartphones and tablet computers. Its principal purpose is as part of the spell checker to correct common spelling or typing errors, saving time for the user. It is also used to automatically format text or insert special characters by recognizing particular character usage, saving the user from having to use more tedious functions.

The replacement list for text replacement can also be modified by the user, allowing the user to use shortcuts. If, for example, the user is writing an essay on the industrial revolution, a replacement key can be set up to replace "ir" with "industrial revolution", saving the user time whenever they want to type it. For users with the patience, this facility can even be used to create a complete keyboard shorthand system, along lines similar to those of Dutton Speedwords, but with short forms instantly replaced by full forms.

POINT BASED REPUTATION:

Reputation systems are programs that allow users to rate each other in online communities in order to build trust through reputation. Some common uses of these systems can be found on E-commerce websites such as eBay, Amazon.com, and Etsy as well as online advice communities such as Stack Exchange. These reputation systems represent a significant trend in "decision support for Internet mediated service provisions".[1] With the popularity of online communities for shopping, advice, and exchange of other important information, reputation systems are becoming vitally important to the online experience. The idea of reputations systems is that even if the consumer can't physically try a product or service, or see the person providing information, that they can be confident in the outcome of the exchange through trust built by recommender systems.

B) Categories and tags using LDA Algorithm:

LDA is a generative probabilistic model of a corpus. The basic idea is that the documents are represented as random mixtures over latent topics, where a topic is characterized by a distribution over words. Latent Dirichlet allocation (LDA), first introduced by Blei, Ng and Jordan in 2003(Blei et al., 2003), is one of the most popular methods in topic modeling. LDA represents topics by word probabilities. The words with highest probabilities in each topic usually give a good idea of what the topic is can word probabilities from LDA.

LDA, an unsupervised generative probabilistic method for modeling a corpus, is the most commonly used topic modeling method. LDA assumes that each document can be represented as a probabilistic distribution over latent topics, and that topic distribution in all documents share a common Dirichlet prior. Each latent topic in the LDA model is also represented as a probabilistic distribution over words and the word distributions of topics share a common Dirichlet prior as well. Given a corpus D consisting of M documents, with document d having N d words (d ∈{1,..., M}), LDA models D according to the following generative process (a)Choose a multinomial distribution φ t for topic t (t ∈{1,..., T}) from a Dirichlet distribution with parameter β.

(b) Choose a multinomial distribution θ d for document d (d ∈{1,..., M}) from a Dirichlet distribution with parameter α.

(c)For a word wn (n ∈{1,..., Nd }) in document d,

(i) Select a topic z n from θ d .

(ii) Select a word wn from φ zn .

In above generative process, words in documents are the only observed variables while others are latent variables (φ and θ) and hyper parameters (α and β). In order to infer the latent variables and hyper parameters, the probability of observed data D is computed and maximized as follows:

LDA is a distinguished tool for latent topic distribution for a large corpus. Therefore, it has the ability to identify sub-topics for a technology area composed of many patents, and represent each of the patents in an array of topic distributions. With LDA, the terms in the collection of documents produce a vocabulary that is then used to generate the latent topics. Documents are treated as a mixture of topics, where a topic is a probability distribution over this set of terms. Each document is then seen as a probability distribution over the set of topics. We can think of the data as coming from a generative process that is defined by the joint probability distribution over what is observed and what is hidden.

C) Email notifications using LIP-KIMAA:

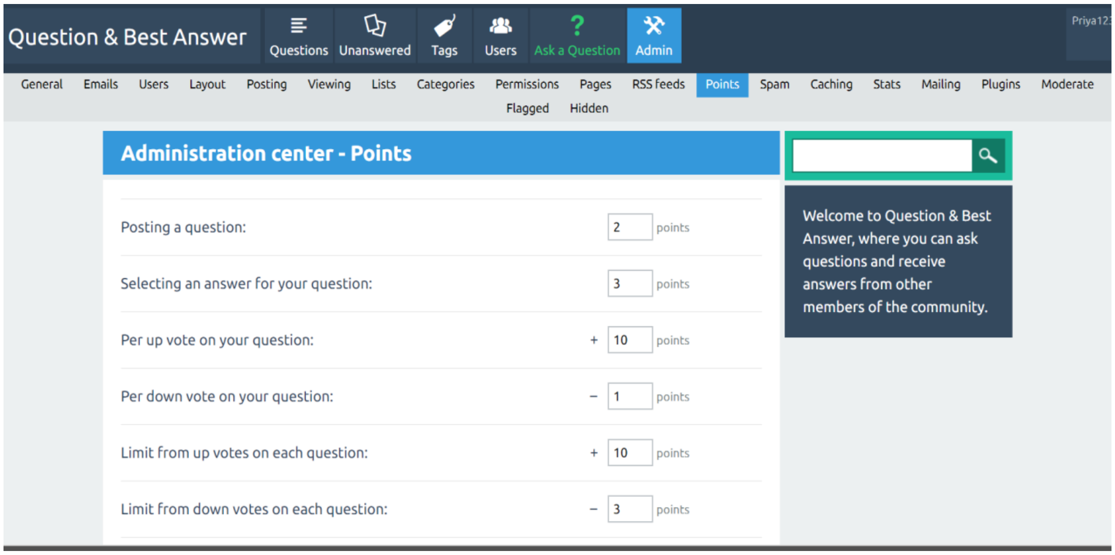
Alert messaging (or alert notification) is machine-to-person communication that is important or time sensitive. An alert may be a calendar reminder or a notification of a new message.Alert messaging emerged from the study of personal information management (PIM), the science of discovering how people perform certain tasks to acquire,organize, maintain, retrieve and use information relevant to them. Alert notification is a natural evolution of the concept of RSS which makes it possible for people to keep up with web sites in an automated manner. Alerting makes it possible for people to keep up with the information that matters most to them.Alerts are typically delivered through a notification system and the most common application of the service is machine-to-person communication. Very basic services provide notification services via email or SMS. More advanced systems (for example AOL) provides users with the choice of selecting a preferred delivery channel such as e-mail, Short Message Service (SMS), instant messaging (IM), via voice through voice portals, desktop alerts and more. Novel approaches provide users with the ability to schedule their own alerts (for example Google Calendar). The most sophisticated service providers embrace all capabilities, aggregating a multitude of reminder, notifications and alert, catering the delivery system to the specific context of the content being delivered thus enabling users to create sophisticated scenarios.

D) Voting, comments and follow-on questions using LIP-IMF algorithm:

The proposed LIP-KIM and its two approximate algorithms (LIP-KIMA and LIP-LIMAA) are comprehensive. In terms of the modeling power, they capture all the three aspects (nonlinearity, coupling, and dynamics). In this section, we show that our algorithms are also flexible. For example, if only a subset of these three aspects matter with the prediction performance for some applications, our algorithms can be naturally adapted to these special cases (e.g., see LIP-IM and LIP-KIA in the supplementary material, online available). Here, we further show the flexibility of our models in terms of temporal forgetting and feature/example selection

VII. OUTPUT

STEP 1:Points-based reputation management & auto correction on question using KIMA algorithm:

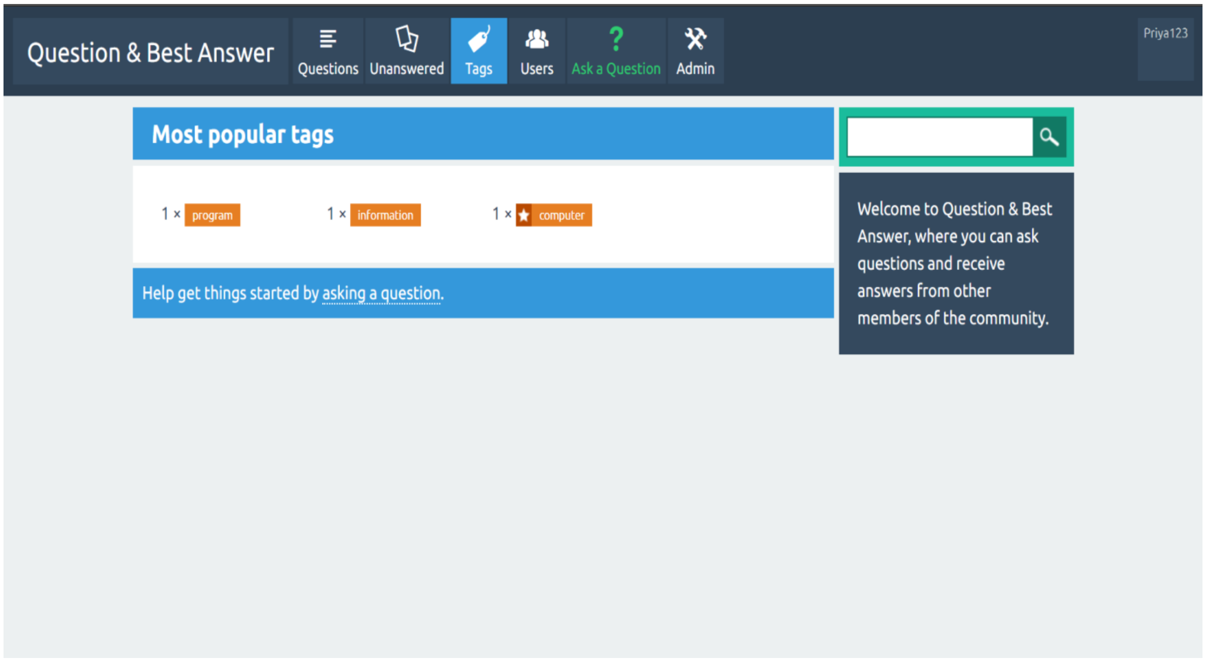


Administration center - Points

Posting a question:	2	points
Selecting an answer for your question:	3	points
Per up vote on your question:	+ 10	points
Per down vote on your question:	– 1	points
Limit from up votes on each question:	+ 10	points
Limit from down votes on each question:	– 3	points

Welcome to Question & Best Answer, where you can ask questions and receive answers from other members of the community.

STEP 2: Categories and tags using LDA Algorithm

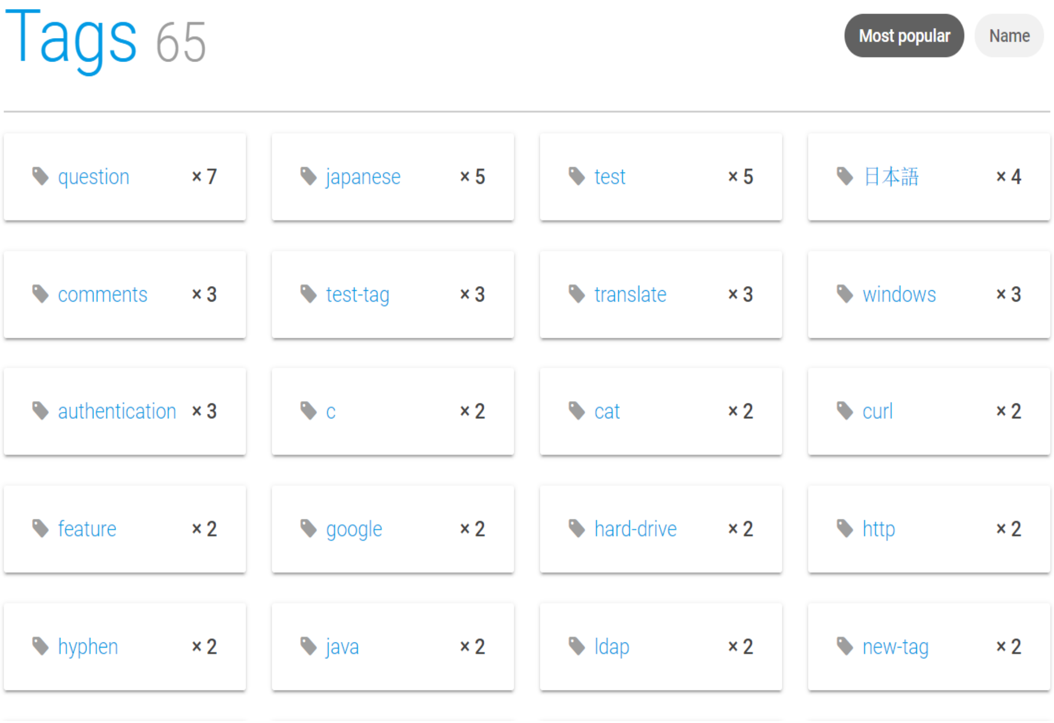


Most popular tags

1 × program 1 × information 1 × ★ computer

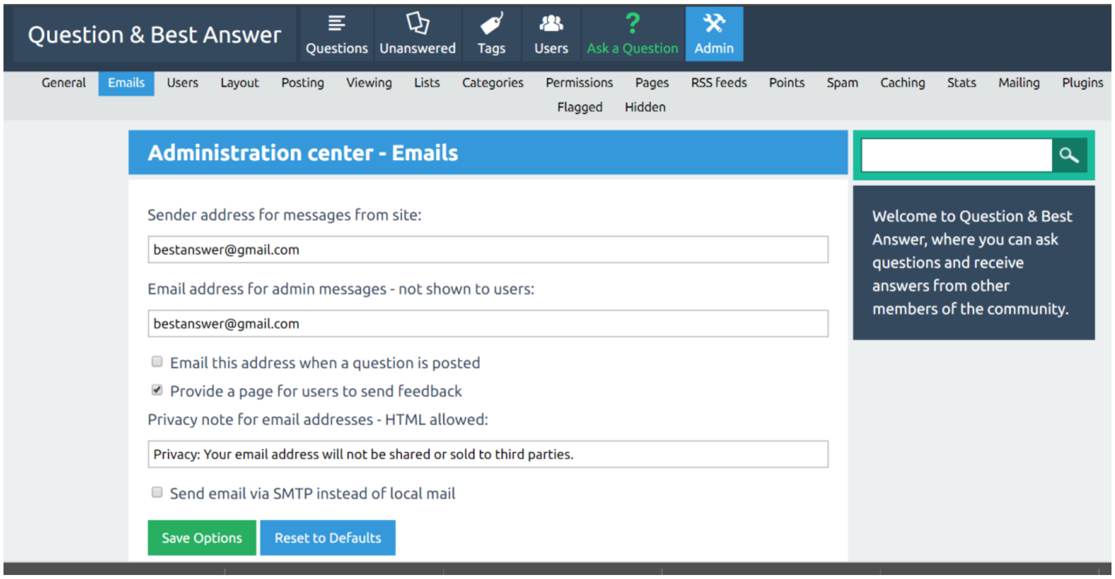
Help get things started by asking a question.

Welcome to Question & Best Answer, where you can ask questions and receive answers from other members of the community.



question × 7	japanese × 5	test × 5	日本語 × 4
comments × 3	test-tag × 3	translate × 3	windows × 3
authentication × 3	c × 2	cat × 2	curl × 2
feature × 2	google × 2	hard-drive × 2	http × 2
hyphen × 2	java × 2	ldap × 2	new-tag × 2

STEP 3: Email notifications using LIP-KIMAA



Administration center - Emails

Sender address for messages from site:

bestanswer@gmail.com

Email address for admin messages - not shown to users:

bestanswer@gmail.com

☐ Email this address when a question is posted

☑ Provide a page for users to send feedback

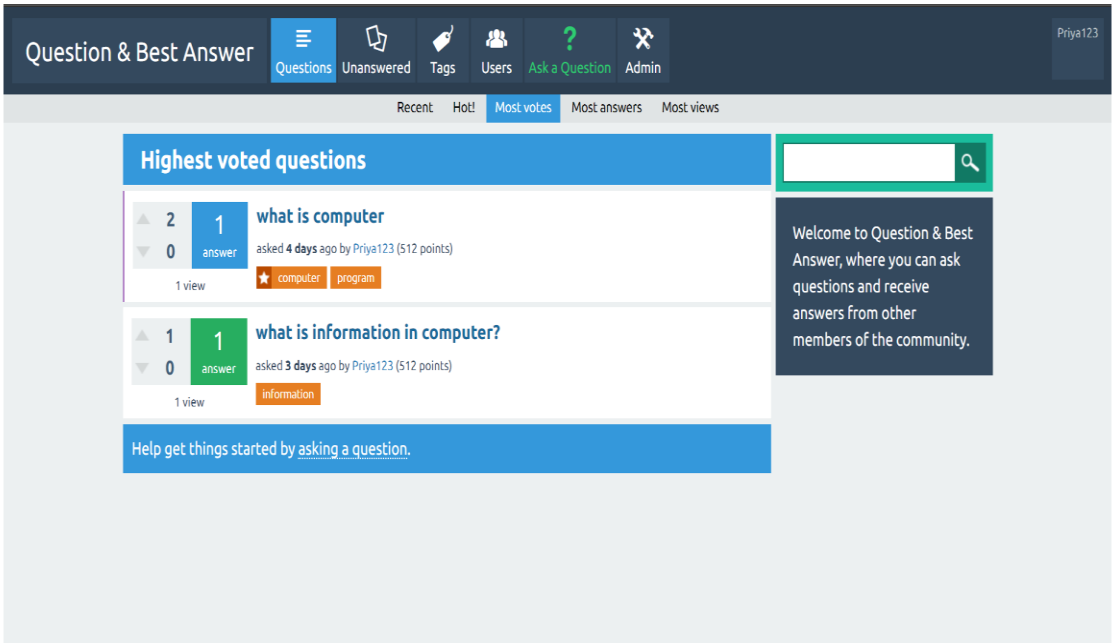
Privacy note for email addresses - HTML allowed:

Privacy: Your email address will not be shared or sold to third parties.

☐ Send email via SMTP instead of local mail

Welcome to Question & Best Answer, where you can ask questions and receive answers from other members of the community.

STEP 4: Voting, comments and follow-on questions using LIP-IMF algorithm.



Highest voted questions

2 0 1 answer 1 view — what is computer — asked 4 days ago by Priya123 (512 points) — computer, program

1 0 1 answer 1 view — what is information in computer? — asked 3 days ago by Priya123 (512 points) — information

Help get things started by asking a question.

Welcome to Question & Best Answer, where you can ask questions and receive answers from other members of the community.

VIII.CONCLUSION:

we have proposed a family of algorithms to comprehensively and efficiently predict the voting scores of questions/answers in CQA sites. In particular, some of the proposed algorithms (LDA, LIP-KIMA, and LIP-KIMAA) can capture three key aspects that matter with the voting score of a post, while others can handle the special cases when only a fraction of the three aspects are prominent. In terms of computation efficiency, some algorithms (LIP-IM, LIP-IMF, LIP-KIA, LIP-KIMAA, and LIP-KIMAA) enjoy linear, sub-linear,or even constant scalability. The proposed algorithms are also able to fade the effects of old examples (LIP-KIM), and select a subset of features/examples (LIP-MS and LIP-KMS). We analyze our algorithms in terms of optimality, correct-ness, and complexity, and reveal the intrinsic relationships among different algorithms. We conduct extensive experimental evaluations on two real data sets to demonstrate the effectiveness and efficiency of our approaches.

IX.FUTURE ENHANCEMENT:

In our project we have done the best search for question in wamp server .In future we can implement the project in web servers.here we have developed a project in website in future it can be implemented in mobile applications for android and IOS devices.

X. REFERENCES

[1] Agichtein .E, Castillo .C, Donato .D, Gionis .A, and Mishne .G, “Finding high-quality content in social media,” in Proc. 2nd ACM Int. Conf. Web Search Data Mining, 2008, pp. 183–194.

[2] Anderson .A, Huttenlocher .D, Kleinberg .J, and Leskovec .J, “Discovering value from community activity on focused question answering sites: A case study of stack overflow,” in Proc. 18th Int. Conf. Knowl. Discovery Data Mining, 2012, pp. 850–858.

[3]Chang .C and Lin C.-J., “LIBSVM: A library for support vector machines,” ACM Trans. Intell. Syst. Technol., vol. 2, no. 3, p. 27, 2011.

[4] Chen .X, Lin .Q, Kim .S, Carbonell .J.G and Xing.E.P,“Smoothing proximal gradient method for general structured sparse regression,” Ann. Appl. Statist., vol. 6, pp. 719–752, 2012.

[5] Horowitz .D and Kamvar .S, “The anatomy of a large-scale social search engine,” in Proc. 19th Int. Conf. World Wide Web, 2010,pp. 431–440.

[6] Jeon .J, Croft .W, Lee . J, and Park .S, “A framework to predict the quality of answers with non-textual features,” in Proc. 29th Annu
Int. Conf. Res. Develop. Inf. Retrieval, 2006, pp. 228–235.

[7] Jiuming Huang, Hongmei Liu, Xiaolei Fu, Chao An” Answer Quality Prediction Joint Textual and Non-Textual Features” IEEE transactions on Web Information Systems and Applications

[8]Pan .B , et .al , “Incremental kernel ridge regression for the prediction of soft tissue deformations,” in Medical Image Computing and Computer-Assisted Intervention. Berlin, Germany: Springer, 2012,pp. 99–106.

[9] Piyush Arora, Debasis Ganguly and Gareth Jones .J.F” The Good, the Bad and their Kins: Identifying Questions with Negative Scores in StackOverflow” ieee transactions on Advances in Social Networks Analysis and Mining

[10] Shah . C and Pomerantz .J, “Evaluating and predicting answer quality in community QA,” in Proc. 32rd Annu. Int. Conf. Res. Develop. Inf. Retrieval, 2010, pp. 411–418.