International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 11, ISSN No. 2455-2143, Pages 131-137
Published Online March 2023 in IJEAST (http://www.ijeast.com)

FINDING RELATED FORUM POSTS THROUGH TWITTER DATA USING MATCHING ALGORITHM

Dr.K.N.S.Lakshmi, B.Madhu Sowmya
Department of CSE
SVPEC, Visakhapatnam, A.P, India

Abstract— **Forum posts has the specific problem of finding related posts to a post at hand. By considering across the related documents the contents of posts are generally consider are whole. Here similarity process are done between two posts with respective segments and should be of same intention. All posts are generally fragmented in the form of group to attain the goal bunches. Now similarities are generally cross view the forums with sections and that will be of same intention.**

Keywords— **Segmentation**

I. INTRODUCTION

Forums are generally a online discussion site, where people hold there conversations by posts. It is like a message board and different from chat rooms. A traditional approach for finding related document that perform content comparisons across content of posts, the contents are compared by different posts. The relatedness of two posts can then be based on a comparison across segments that serve the same goal. Every posts are generally considered as segments. Segments are generally said as parts (or) sections. In This the relatedness between two posts should be based on similarities respective to segments. The segmentation methods play important role by developing work with monitoring the no of text features ,it identify by parts of post. While this process performing significant jumps are occurred because of that segmentation are occur. Now segmentation of all posts are generally clustered in the form of intention cluster so that the similarities are calculated across segmentation with same intention. plays important role.

1) DATA MINING:

. Data mining also called as knowledge Discovery refer to extracting interesting patterns and useful knowledge from huge amount of data. Data mining in the databases is a new interdisciplinary field, merging ideas from statistics, machine learning, databases and parallel computing. Knowledge Discovery in Database(KDD) is a process of finding useful information and patterns in data where as Data mining is the use of algorithms to extract the information and patterns derived by the KDD process.

CLUSTERING:

Clustering in data mining is a discovery process that groups a set of data. The principle of clustering is "Maximizing the Intra
cluster similarity and minimizing the Inter cluster similarity." That is, a cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in the other clusters. A cluster of data can be treated collectively as one group in many applications.

SEGMENTATION

Segmentation is a key data mining technique. The key to segmentation is to decide how to split the database up. Segment is a group of consumers that react in a similar way to a particular marketing approach. So the key to segmentation is to decide how to split the database up.

II. EXISTING SYSTEM

There is no goal base division in Existing System. Question Answering Communities (QAC) plays important role they coordinate the work for posts. Division techniques are isolated into 2 general gatherings. The first is topical division where contiguous sets of content pieces are thought about for general likeness in light of terms or points. The second gathering of division strategies comprises of Transcribed oral-talk procedures utilized as a part of the investigation of translated oral correspondence utilizing semantic criteria.

1.3.1. Advantages of Existing System

- No complexity in computations.

1.3.2.Disadvantages of Existing System

- Less Efficiency. .Complexity of TF/IDF algorithm is high.
- Difficult to finding related forum posts
- Segmentation increases cost
- Searching is difficult
- Complexity of TF/IDF algorithm is high

International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 11, ISSN No. 2455-2143, Pages 131-137
Published Online March 2023 in IJEAST (http://www.ijeast.com)

III. PROPOSED SYSTEM:

3.1.SEGMENTATION OF POSTS

The challenging task is finding right segmentation process, from segmented documents, that are occurred in large body of work, if it is in the form of document d, there are 2^{jdj1} possible segmentations are occur.

Every segmentation process should be in the form of

- coherent
- largely disconnected from its adjacent segments.

Segmentation is the intention-based, these two properties translate to a segmentation where every segment conveys a single clear intention conveys by the adjacent segments.

3.2. SEGMENT GROUPING

Segments with similar intentions are created same group and segments with different intentions in different groups. It is modeled with vector of features, array of information are taken here. Now each cluster are generally communicates respectively the same goal.I to denote a cluster, and C to denote the set of the generated clusters.

Vector of weights that are based on the feature values are created by us. Vector with the letter F. Now consider two types of weights that capture the strength of the use of each CM categorical value, of each feature.

Now each CM value with in the segment are measured then the comparisons are done to categorical values that belong to same communication appearing in segments.Using the notion of the distribution table DSb_{CMr} of a communication mean CMr introduced in Section we define the vector F_s of weights, one weight for each feature.

3.3.SEGMENTATION REFINEMENT:

They should have same document with same segments that are end up with same cluster with same intention., if they have the same intention but are not same cluster then consequence document. The segments that belong to the same document in a cluster are concatenated into one.

Advantages of Proposed System

- Complexity is low.
- Searching became easy.
- Finding related forum posts became more efficiency.
- Searching became now easy

Ⅳ. ALGORITHM:

3.4.1.Matchin Algorithm:

Document matching is one of the best technique plays important role by collecting of documents that are generally related to reference document dq .Now the dq reference document are measure the relatedness between other documents d0 are lie in the form of IR technique.

3.4.1.1.Matching with respect to a specific Intention:

Every document with some specific intention are projected on each cluster. The specific intention are made by measuring the related documents to reference document d0. Text comparison are computed the relatedness between the documents like IR technique i.e. TF/IDF model.TF/IDF method and its probabilistic variance BM25 consists of a term weighting scheme that weighs a term in a document considering the number searching That variance computes the weight of a termt in a document d^0. If s_q and s^0 are the segments of the documents d_q and d^0, respectively, in the intention cluster. where fsq (t) denotes the frequency of the term t in the segment sq, jIj the cardinality of the intention cluster, and jI^tj the number of segments in the intention cluster.

3.4.1.2. Matching with respect to All the Intentions:.

This algorithm consist of top-nlists generated across the different intentions, ., the set M are used to generate the k most related documents to the reference document d_q R is created as new list contains in every document in lists in M. Each document are associated with the sum of the scores with which this document appears in the various lists in M. The k elements in R with the highest score are returned as answer to the request of the matching documents to the reference document d_q.High value for n compared to the value of k, on the other hand, will favor documents that appear in many lists even with not very high scores. We have empirically found that a good choice is an n equal to 2 k

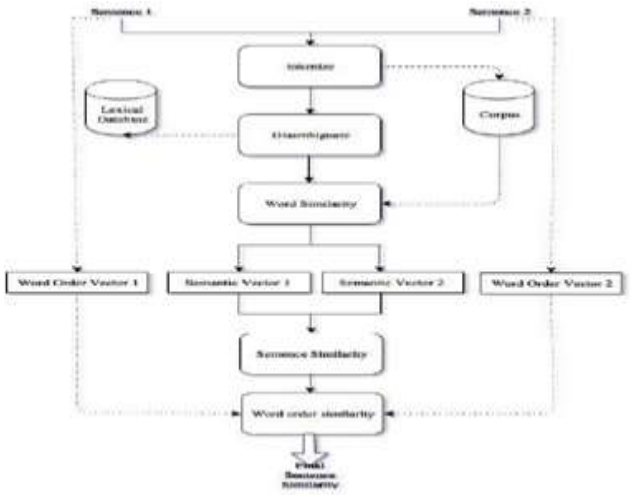
3.4.2. Clustering Algorithm

Clustering is the task of dividing the data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar trait s and assign them into cluster. It is the most important unsupervised learning problem

K-Mean Method

If a vector of documents (Dl, D2 . . · Dn) is given ,K-means clustering Algorithm will partition the n documents into K clusters (K :::; n) such that cosine distance between them is minimum.

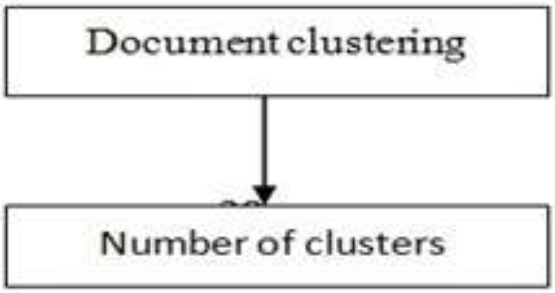
International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 11, ISSN No. 2455-2143, Pages 131-137
Published Online March 2023 in IJEAST (http://www.ijeast.com)



1. Randomly select initial centroid that divides the documents into k clusters.
2. Compute Cosine distance of each document from the centroid of each of the clusters. Assign that document to the cluster with the closest centroid.
3. Repeat step 2 until there is no change in newly formed clusters.

$$CS\ ixj = \begin{pmatrix} 1 & D(1,2) & D(1,3) & D(1,4) \ldots D(1,n) \\ D(2,1) & 1 & D(2,3) & D(2,4) \ldots D(2,n) \\ D(3,1) & D(3,2) & 1 & D(3,4) \ldots D(3,n) \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ D(n,1) & D(n,2) & D(n,3) & D(n,4) \ldots D(n,n) \end{pmatrix}$$

3.4.3. The Proposed Method As Follows:



3.4.3. 1 TF/IDF Algorithm

It is generally said as term frequency–inverse document frequency. It is often used as a weighting factor in searches of information retrieval, text mining, and user modeling. TF-IDF, a vector space based representation is the common technique used for text processing.

TF–IDF weighting scheme are often used for in scoring and ranking a document's relevance given a user query. TF–IDF can be successfully used for stop-words filtering in various subject fields, including

- Text Summarization
- Classification

The TF–IDF IS calculated as:

TF-> Term Frequency

IDF->Inverse Document Frequency

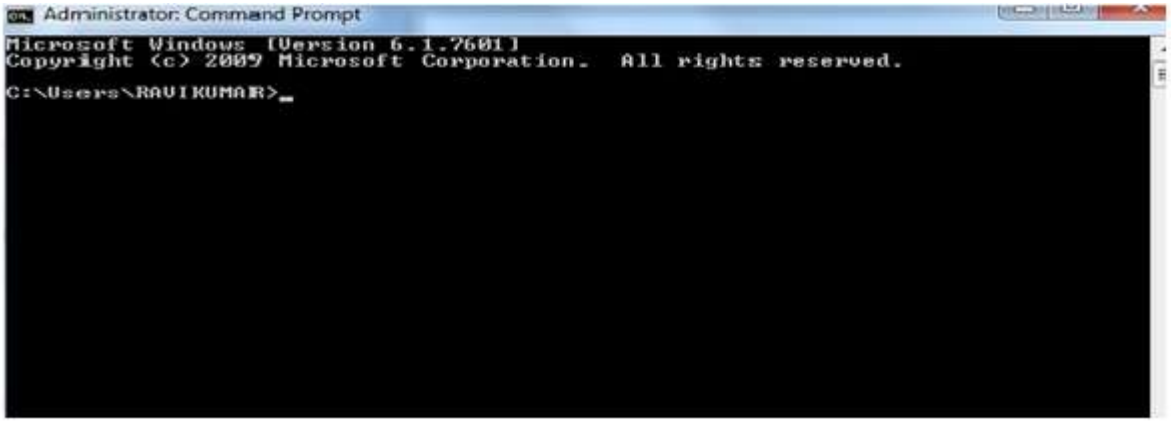
$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \cdot \mathrm{idf}(t, D)$$

where
t=term
d=referencing document
D=set of Documents

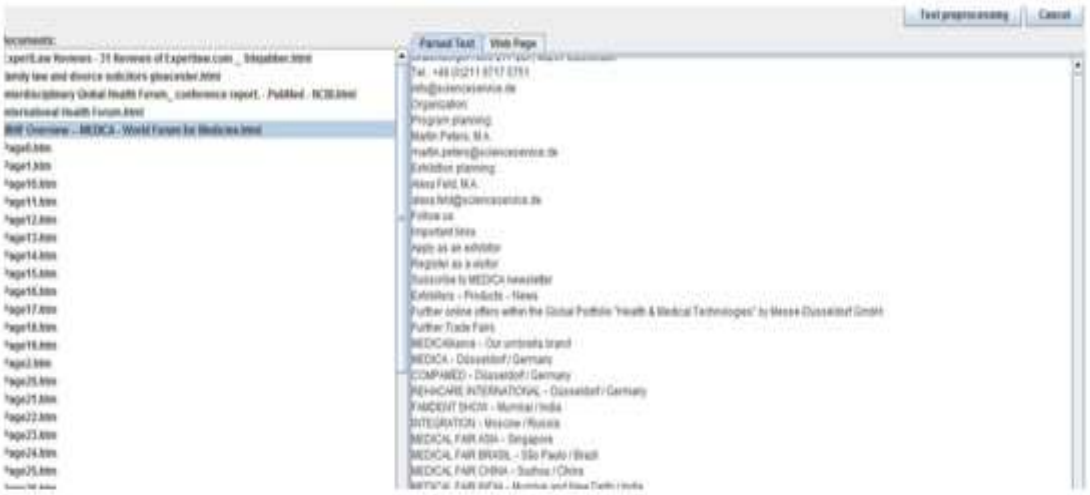
International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 11, ISSN No. 2455-2143, Pages 131-137
Published Online March 2023 in IJEAST (http://www.ijeast.com)

IV. RESULTS AND DISCUSSION:

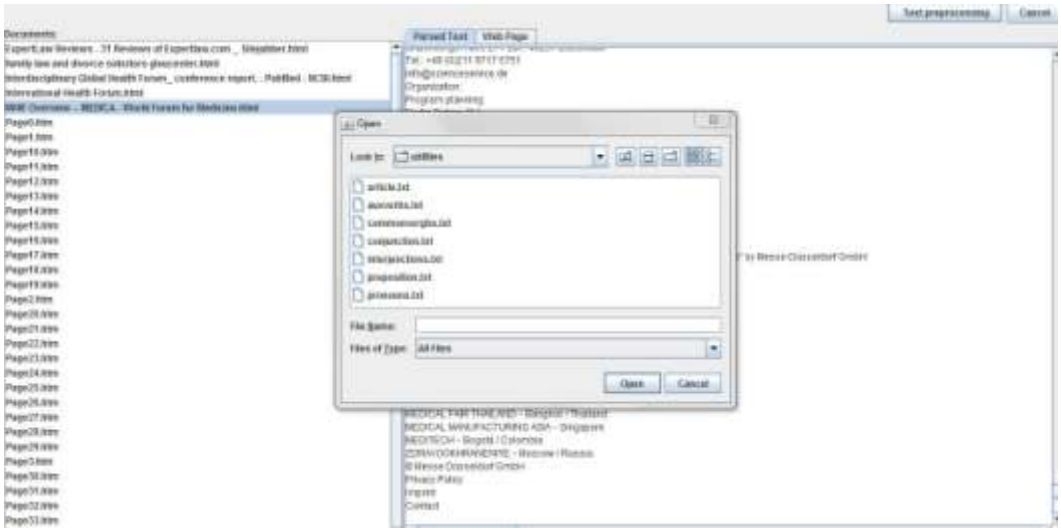
i. OPEN THE COMMAND PROMPT AND RUN



ii. IT IS THE FIRST PAGE

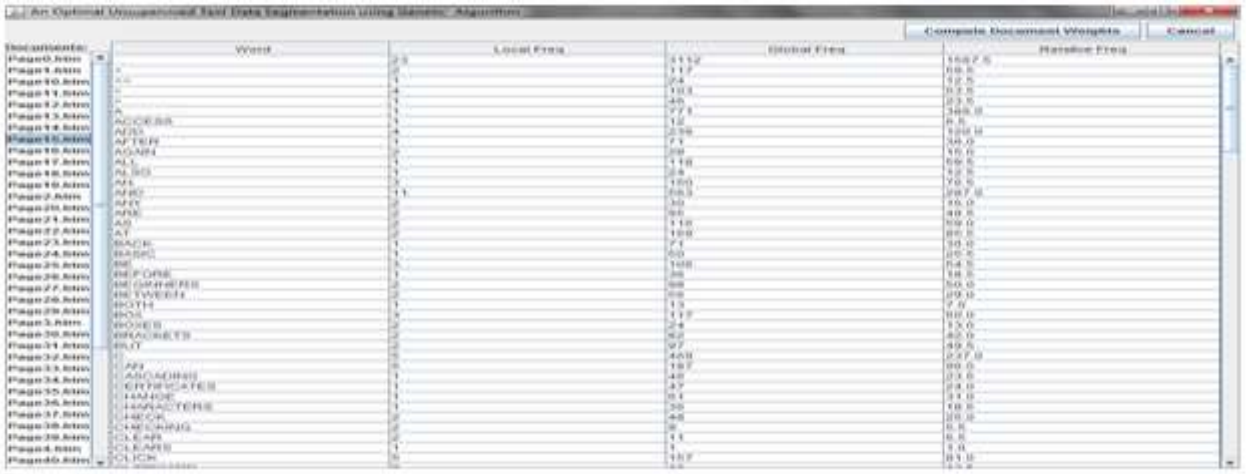


iii .NOW PRE PROCESSING IS TO BE PERFORMED

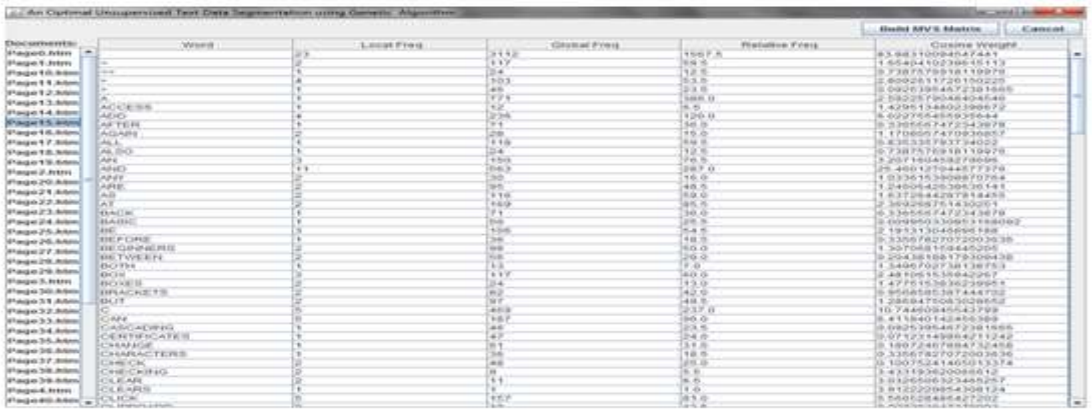


International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 11, ISSN No. 2455-2143, Pages 131-137
Published Online March 2023 in IJEAST (http://www.ijeast.com)

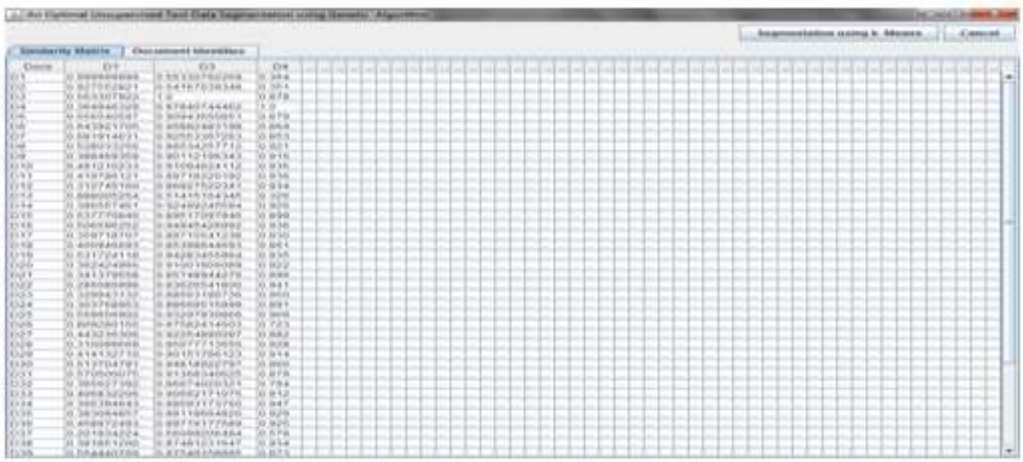
iv.RESULT AFTER PREPROCESSING.



v.COMPUTE DOCUMENT WEIGHTS



VI. BUILT VSM MATRIX



International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 11, ISSN No. 2455-2143, Pages 131-137
Published Online March 2023 in IJEAST (http://www.ijeast.com)

V. OUTPUT

VI. CONCLUSION

This paper presents an approach to calculate the semantic similarity between two words, sentences or paragraphs. A novel approach proposed by us for matching a reference post to the k most related posts in a collection. In our method , segmentation are done across the posts that convey similar with some intentions. We presented several experiments regarding the right segmentation criteria, the effectiveness of the segmentation algorithms and the formation of intention clusters that prove that a rather intuitive concept, that of the author intentions to communicate a certain message, can be effectively captured by an automated process.Then extractive summarization is used to extract feature terms and the summarized sentences were ranked based on feature frequency of the posts, measuring the relatedness score after having distinguished the different

VII. FUTURE SCOPE

- As we are just transferring only the text messages, we can also transfer the files, images etc.
- As we are using MUTUAL AUTHANTICATION PROTOCOL for the communication we can also use OPTIMALITY OF KEYING PROTOCOL.
- In this project we have been used symmetric keys for the communication we can also use ID of the sensors for the communication.

REFERENCES

[1]. M. Chen, X. Jin, and D. Shen, "Short text classification improved by learning multi-granularity topics," in IJCAI, 2011, pp. 1776–1781.

[2]. J. Jeon, W. B. Croft, and J. H. Lee, "Finding semantically similar questions based on their answers," in Proceedings of the 28th ACM SIGIR Conference, ser. SIGIR '05. New York, NY, USA: ACM, 2005, pp. 617–618

[3]. S. Robertson, S.Walker, and M. Hancock-Beaulieu, "Okapi at TREC-7: Automatic ad hoc, filtering, VLC and interactive track," TREC '98, pp. 199–210, 1998

[4]. Jiawei Han, Micheline Kamber, Data Mining and Concepts.

[5]. Maqgret H. Dunham, Data Mining and Introductory to Advanced Topics

[6]. .ClusterAnalysis:- http://www.tutorialspoint.com/data_mining/dm_cluster_analysis.htm

[7]. DataMining: http://www.webdocs.cs.ualberta.ca/~zaiane/courses/comput690

[8]. PorterStemmerAlgorithm: http://snowball.tartarus.org/porter/stemmer.htm

[9]. TF-IDF: http://www.tfidf.com

[10]. K-Means-Algorithm:http://sites.google.com/site/dataclusteringalgorithms/k-means-clustering- algorithm

[11]. SentimentAnalysis – Lexalytics: https://www.lexalytics.com/technology/sentiment

[12]. G. Salton, A. Singhal, C. Buckley, and M. Mitra, "Automatic text decomposition using text segments and text themes," in ACM Hypertext, 1996, pp. 53–65

[13]. S. Louvigne, N. Rubens, F. Anma, and T. Okamoto, "Utilizing social media for goal setting based on observational learning," in ICALT, 2012, pp. 736–737.

[14]. K.Wang, Z. Ming, and T. Chua, "A syntactic tree matching approach to find similar questions in community QAservices," in ACM SIGIR, 2009, pp. 187 – 194.

[15]. A. Shtok, G. Dror, and Y. Maarek, "Learning from the past: Answering new questions with past answers," in WWW, 2012, pp. 759–768.

[16]. K. Jones, C. Van Rijsbergen, B. L. Research, and D. Department, Report on the Need for and Provision of an Ideal Information Retrieval Test Collection, ser. British Library Research and Development reports, 1975.

[17]. J. Kekalainen, "Binary and graded relevance in ir," Inf. Processing & Management, vol. 41, no. 5, pp. 1019 – 1033,2005.

[18]. Z.-Y. Ming, T.-S. Chua, and G. Cong, "Exploring domain specific term weight in archived question search," in Proceedings of the 19th ACM CIKM, ser. CIKM '10. New York, NY, USA: ACM, 2010, pp. 1605–1608.

[19]. H. Wen, W. Zhongyuan, W. Haixun, Z. Kai, and Z. Xiaofang, "Short text understanding through lexical-semantic analysis," in IEEE ICDE, 2015

[20]. J. Jeon, W. B. Croft, and J. H. Lee, "Finding semantically similar questions based on their answers," in Proceedings of the 28th ACM SIGIR Conference, ser. SIGIR '05. New York, NY, USA: ACM, 2005, pp. 617–618.