International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 03 | Mar-2018 www.irjet.net p-ISSN: 2395-0072

Finding Related Forum Posts through Intention-Based Segmentation

Bijinepalli Sai Lakshmi Mounika[1], Chebiyyam Sai Pratyusha[2], Davuluri Sai Meghana[3]

[1,2,3] *Student, Dept. of Computer Science & Engineering, VVIT, Nambur.*

Abstract – *Here the issue of finding related discussion presents on a post at a hand will be considered. In conventional approach content examination will be performed for finding related records, while in this approach we consider each post as an arrangement of sections, each composed because of an alternate objective. Relatedness between the posts ought to be founded on likeness of portions that are passing on a similar objective. This is conceivable just when similar terms weigh diversely in relatedness score in view of section aim in which they are found. In this division technique, by checking number of content highlights the parts of a post can be distinguished and critical hops happen demonstrating purpose of division. Goal bunches will be shaped by grouping the fragments of all posts and after that similitude's will be ascertained crosswise over sections with same expectation. The effectiveness of our division strategy in finding related gathering posts is delineated*.

***Key Words*: Segmentation, Intention, Forums, Posts.**

1. INTRODUCTION

A type of blog that lets users publishes short text updates. Bloggers can usually use a number of services for the updates including instant messaging, e-mail, and twitter and so on. The posts are called micro posts, while the act of using these services to update your blog is called micro blogging.

Posts:

A comment posted by the users for a particular online discussion or forum.

Forum:

An Internet **forum**, or message board, is an online discussion site where people can hold conversations in the form of posted messages.

Gatherings with regards to online client groups offer clients the capacity to look for arrangements and settle on choices in regards to different issues by misusing other clients' involvement. They likewise offer organizations the capacity to interface and bolster their client base. Existing discussions go from areas like wellbeing, law and innovation. The association of the gathering posts into classifications is a component that encourages clients to distinguish all the more effectively those presents related on a theme. In any case, since perusing a substantial number of posts is baffling and tedious, most gathering destinations offer catch phrase seek capacities. However, catchphrase pursuit may not bring about a total arrangement of related posts since the determination of the correct watchwords isn't generally direct. Relatedness has generally been converted into content comparability. Content comparability registered specifically crosswise over gathering posts is, sadly, not exceptionally viable for this situation on the grounds that inquiries are done under particular topical classifications, e.g., printers, or lodgings in New York, in which the substance of the considerable number of posts is in any case comparative. We advocate that when we are estimating the relatedness of two discussion posts, regarding them as composite questions rather than solid elements can prompt more compelling examinations. The gathering posts comprise of various parts each present with an alternate objective. For example, a section may serve to portray an issue that the writer has, another to give foundation data keeping in mind the end goal to put the peruse into setting, a third to express a want, and a fourth to achieve a conclusion. We allude to these parts of a discussion post as sections. The relatedness of two posts would then be able to be founded on an examination crosswise over portions that serve a similar objective, i.e., they are proposed for a similar reason, rather than a correlation of the two posts as wholes. The correlation among content fragments with a similar expectation can be performed by simple searching techniques and SQL queries. In any case, in our approach, given the distinctive aims of the gathering post creator, the significance and significance of a term is assessed in view of the section in which the term is found. Distinctive weights for similar terms have been utilized crosswise over various topical discussion classes or spaces. To the best of our insight, it is the first occasion when that a weighting plan may relegate distinctive weights to a term in posts of the same topical classification; or even inside a similar post.

Distinguishing the sections in a discussion post is a testing undertaking. Discussion posts are commonly maybe a couple passages long, with finish sentences. They don't take after the contracted style utilized as a part of smaller scale sites, however in the meantime, since they are expected for intelligent discourses, they are not verbose and they do not have the auxiliary develops (e.g., segments) ordinarily utilized as a part of full-content records to distinguish topical units.

We have built up a structure for finding related gathering post that depends on the above thought. By misusing the correspondence implies, the framework distinguishes the distinctive portions inside every gathering post and parts the discussion post into these sections. Portions serving a similar aim are distinguished and assembled together. Given a gathering post within reach, its sections are distinguished

and the coordinating score of each fragment with other discussion posts' portions that have a similar goal is figured.

The intentions of the posts are classified as positive and negative based on the comments provided by the users. When the users post a comment the posts are divided based on the set of positive and negative words provided in the algorithm. The algorithm has the flexibility of adding any number of words to it based on the users interest and intention.

- We formally present a novel technique for finding related discussion posts that regards each post as an arrangement of portions and registers content likeness just crosswise over fragments of a similar expectation.
- We give broad investigations genuine clients that affirm the presence of such sections in discussion posts of various spaces, and confirm the viability of the individual advances and choices of our philosophy, including the fringe determination systems, the choice of highlights, and to wrap things up the capacities and weights for catching content element variety.
- We assess the viability of the general approach on the suggestion of related gathering posts utilizing evaluations and criticism by clients in 3 unique spaces.

Data Mining:

Data Mining is one of the popular techniques which are mainly used for collecting data from different perspectives and analyzing it into useful information. The information which can be used to increase revenue and cost. Data mining is software which is analyzing data from different analytical tools. Finding correlations or patterns among different fields in large relational databases and gathering the knowledge from data mining techniques.

Segmentation:

Data mining is the process of extracting previously unknown and actionable information from large, complex databases. Segmentation is a key data mining technique. A 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.

Clustering:

Clustering is the process of using machine learning and algorithms to identify how different types of data are related and creating new segments based on those relationships. Clustering finds the relationship between data points so they can be segmented.

SQL:

SQL stands for Structured Query Language. It includes simple queries to extract the information from large volumes of data.

2. SYSTEM ARCHITECTURE

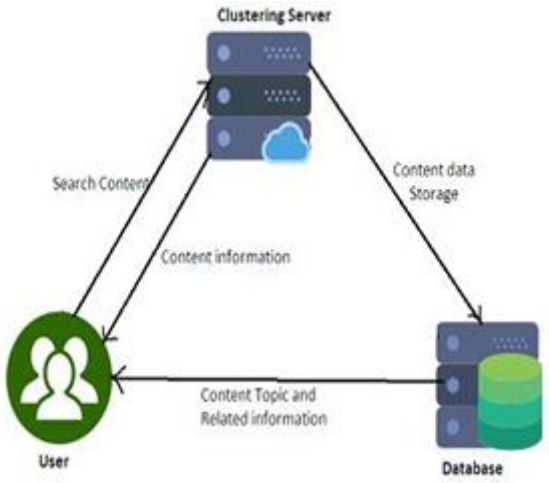

Fig 1: System Architecture

A server cluster is a collection of clusters, called nodes that communicate with each other to make a set of services highly available to clients. If user needs any information, he interacts with clustering server. If related content is found the server directly communicates with the user otherwise server interacts with the database. If the data is found in the database it provides information to the user (Refer Fig 1).

3. EXISTING AND PROPOSED SYSTEM

3.1 EXISTING SYSTEM:

In the Existing System there is no goal based division. There exists significant measure of work for post coordinating in Question Answering Communities (QAC). Individuals look for answers to general intrigue, verifiable or educational, questions.

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.

DISADVANTAGES:

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

3.2 PROPOSED SYSTEM:

The comparison among text segments with the same intention can be performed by Information Retrieval methods, such as one of the many TF/IDF or BM25 variants or language model based methods, or using topics generated by topic modeling techniques like LDA, paraphrasing techniques or even auxiliary external services, with the latter been used especially for documents with short and poor content, e.g., tweets. However in our approach, given the different intentions of the forum post author, the meaning and importance of a term is estimated based on the segment in which the term is found. Different weights for the same terms have been used across different thematic forum categories or domains. To the best of our knowledge, it is the first time that a weighting scheme may assign different weights to a term in posts of the same thematic category; or even within the same post.

The complexity of this algorithm is reduced using simple search queries.

ADVANTAGES:

- Due to the nature of the posts, measuring the relatedness score after having distinguished the different segments/messages that the authors intend to communicate has been proved more effective than the direct comparison of the whole posts.
- Easy to finding related forum posts with more effectively and more efficiently.
- Complexity is low.
- More Efficient

4. IMPLEMENTATION

- The user cannot access his account until the Admin authorize the user status as authorized.
- Admin can see the all the users who are present in the database.
- The user cannot post more than 2 lines in micro blogs.

4.1 Create Forums

Only the Admin can create the forum posts. First, the Admin log into his account and he creates the forum post category and then he gives the unique name to the post and description to the forum post.

4.2 Search Posts

Here, the user first log into his account. He then searches the post using keyword (Searching content based on post description). If the post is present in the database it will be displayed along with post name and domain.

4.3 Intention based Matching

When the user comments on the post the post is divided into set of segments. Based on the segmentation algorithm the posts are classified as positive and negative intention posts (Refer Fig 2).

This positive and negative intention posts are displayed in the admin home page where the admin can see how many positive and negative posts are posted by the users for a particular forum post. Based on the number of positive and negative intentions for a particular post the positive post chart (Refer Fig 5) and negative post chart (Refer Fig 6) are created.

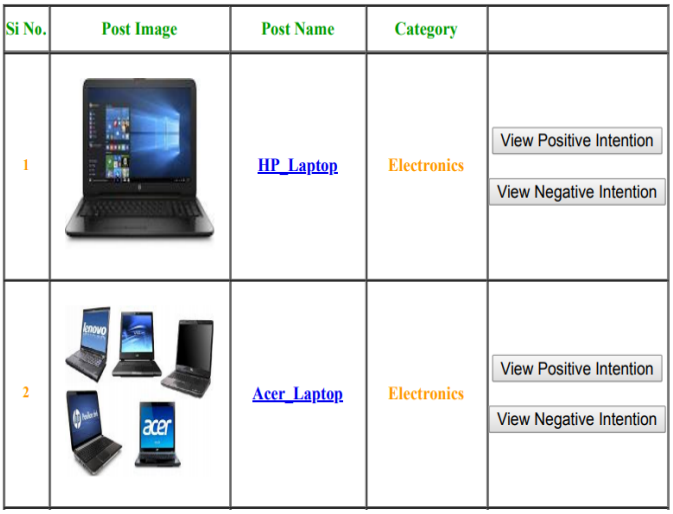

Fig 2: Positive and Negative Intentions

Table 1: Some of the positive and negative words for segmentation

Positive	Negative
Good	Bad
Yes	No Not None
Do Does	Don't Doesn't
Was Were	Wasn't Weren't
Has Have Had	Hasn't Haven' Hadn't
Better Cheap	Abusive Sucks

Based on the words present in the table (Refer Table 1) the posts will be divided as positive or negative. We can add any

number of positive and negative words to the algorithm based on our interest.

4.4 Recommend posts

The user can also recommend posts to his friends if he found the post is useful. This is the additional feature implemented in this paper that to this feature doesn't exist in the micro blogs till now.

4.5 Segmentation grouping

The same intention posts are clustered together into a single cluster. The cluster consists of all the posts that are intended for the same goal/intention.

For ex:

Positive intention posts →in single cluster
Negative intention posts →in single cluster

Positive Discussions on Post HP_Laptop ..

Intention By	Intention Details	Date
Rajesh	It is good laptop	30/12/2017 15:43:58
Rajesh	It is good laptop	30/12/2017 15:44:46

Fig 3: Positive Intention Posts

On clicking the View Positive Intention button the admin is displayed with a page containing a cluster of all the positive comments related to a particular post (Refer Fig 3).

If the admin clicks on the View Negative Intention button the admin is displayed with a page containing a cluster of all the negative comments related to a particular post (Refer Fig 4).

Negative Discussions on Post Dotnet ..

Intention By	Intention Details	Date
Manjunath	It is bad in installation	30/12/2017 16:06:42
Ramesh	It is bad in plug ins	30/12/2017 16:08:01

Fig 4: Negative Intention Posts

Delete Forums

The admin has the right to delete forums if he found more than five bad intentions to the same post. The admin provides the reason of why he was deleting the post. Once the post is deleted by the admin, it will be no longer available in the database.

5. ALGORITHMS

5.1 TF/IDF Algorithm

- It is often used as a weighing factor in searches of Information retrieval.
- It is one of the simplest ranking functions and is computed by summing the TF/IDF query term.
- This algorithm is mainly used for searching process.
- Then 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

The equation has a higher complexity in the existing system and also searching process takes more time. So, we reduced the complexity of this algorithm using simple queries and searching methods.

Example SQL query used in this paper for searching purpose:

String str="select * from allcomments where p_name='"+p_Name"' and categorie='"+p_Categorie"';

5.2 Clustering Algorithm

Clustering is a process which partitions a given data set into homogeneous groups based on given features such that similar objects are kept in a group whereas dissimilar objects are in different groups. It is the most important unsupervised learning problem.

It deals with finding structure in a collection of unlabeled data.

6. SEGMENTATION EVALUATION

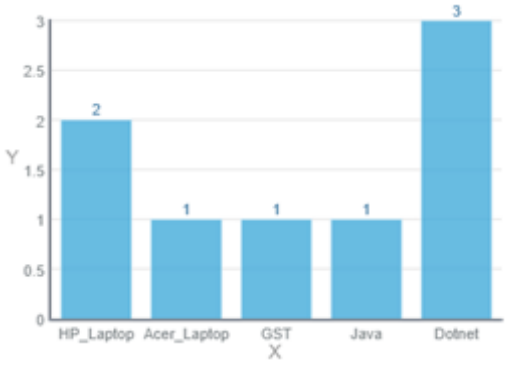

Fig 5: Positive Segmentation chart

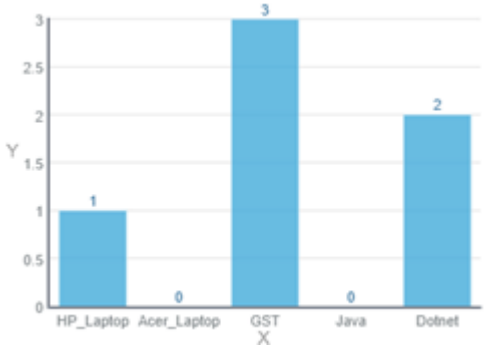

Fig 6: Negative Segmentation chart

Here, the segmentation chart will be formed by considering the information of respective clusters. The positive cluster consists of all the positive posts regarding a topic and the negative cluster consists of all negative posts.

7. CONCLUSION

We proposed a novel approach for matching a reference post to the k most related posts in a collection. Our method identifies and exploits post segments that convey similar author 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. Moreover, due to the nature of the posts, measuring the relatedness score after having distinguished the different segments/messages that the authors intend to communicate has been proved more effective than the direct comparison of the whole posts.

REFERENCES

[1] 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.

[2] T. C. Zhou, C.-Y.Lin, I. King, M. R. Lyu, Y.-I.Song, and Y. Cao, "Learning to suggest questions in online forums." in AAAI, 2011.

[3] L. Weng, Z. Li, R. Cai, Y. Zhang, Y. Zhou, L. T. Yang, and L. Zhang, "Query by document via a decomposition-basedtwo-level retrieval approach." Association for Computing Machinery, Inc., July 2011.

[4] V. Govindaraju and K. Ramanathan, "Similar document search and recommendation," Journal of Emerging Technologies in Web Intelligence, vol. 4, no. 1, pp. 84–93, 2012.

[5] 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.

[6] D. M. Blei, "Probabilistic topic models," Commun. ACM, vol. 55, no. 4, pp. 77–84, Apr. 2012.

[7]J. Berant and P. Liang, "Semantic parsing via paraphrasing." in ACL (1), 2014, pp. 1415–1425.

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

[9] Z.-Y. Ming, T.-S.Chua, and G. Cong, "Exploring domainspecific 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.

[10] D. Papadimitriou, G. Koutrika,Y.Velegrakis and J. Mylopoulos, "Finding Related Forum Posts through Content Similarity over Intention-Based Segmentation" in IEEE Transactions on Knowledge and Data Engineering, vol. 29, no. 9, pp. 1860-1873, Sep 2017.

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

[12] M. Hearst, "Texttiling: Segmenting text into multi-paragraph subtopic passages", Comput. Ling., *vol. 23*, pp. 33-64, 1997.

[13] H. Misra, F. Yvon, J. M. Jose, O. Cappe, "Text segmentation via topic modeling: an analytical study", Proc. Conf. Inf. Knowl. Manage., pp. 1553-1556, 2009.

[14] X. Hu, N. Sun, C. Zhang, T. Chua, "Exploiting internal and external semantics for clustering short texts using world knowledge", Proc. Conf. Inf. Knowl. Manage., pp. 919-928, 2009.

[15] I. Hulpus, C. Hayes, M. Karnstedt, D. Greene, "Unsupervised graph-based topic labelling using dbpedia", Proc. ACM Int. Conf. Web Search Data Mining, pp. 465-474, 2013.