# Finding Related Forum Posts through Content Similarity over Intention-based Segmentation (Extended Abstract)

Dimitra Papadimitriou #[1], Georgia Koutrika *[2], Yannis Velegrakis #[3], John Mylopoulos #[4]

# *University of Trento, Italy,* * *Athena Research Center, Greece*

[1] d.papadimitriou@unitn.it [2] georgia@imis.athena-innovation.gr [3] velgias@disi.unitn.eu [4] jmylopou@uottawa.ca

***Abstract*—We study the problem of finding related forum posts to a post at hand. We developed a multi-segment matching technique that considers posts as a set of segments each one written with a different goal in its author mind and computes the relatedness between two posts based on the similarity of their respective segments that are intended for the same goal. The questions are how our method identifies such segments, how it figures out for what each segment is intended and how it exploits this information to rank the posts. We experimentally illustrate the effectiveness and efficiency of our segmentation method and our overall approach of finding related forum posts.**

## I. Introduction

Finding posts related to a post at hand in a forum in the context of an online user community allows users to seek solutions and make decisions on diverse problems such as health (e.g., *Medhelp*), law (e.g., *ExpertLaw*) and technology problems (e.g., *HP support forum*) based on other user experience without having to formulate complex queries, or perform complicated, long browsing. Work towards this direction has been done for questions in Q&A archives [1], [2] but not for richer-content posts.

To solve the problem, we need to be able to compute some relatedness score, referred to as *matching score*, of every post in a collection to a reference post. Content similarity [3] computed *directly across forum posts* is, unfortunately, not very effective in the case of forums because searches are done under specific thematic categories, e.g., *printers, or hotels in New York*, in which the content of all the posts is anyway similar. In other words, content similarity of the posts as wholes does not necessarily imply relatedness.

Since at the moment of text construction, the author selects the words and the text structure that most effectively fulfill a goal s/he has in mind, we advocate that relatedness can be better assessed by computing a score, not across the content of two posts as a whole, but across those parts, referred to as *segments*, that express the same intention. For instance, a segment may serve to describe a problem that the author has, or to provide background information in order to put the reader into context, to express a desire, or to reach a conclusion. Specifically, we have developed an *unsupervised multi-segment document matching technique* that clusters segments with similar intentions together and provides the top-k forum posts related to a reference post by considering content similarities within each cluster to derive an overall score between each forum post and the reference post. We will use the term post and document interchangeably.

## II. Problem

A document can be seen as a sequence of non-overlapping segments, the concatenation of which is the document itself. Its division into such a sequence is known as *segmentation*. The goal for which a segment has been written may not be explicitly stated, but through the way the segment is constructed, it is reflected into the characteristics of the text. Thus, monitoring and identifying strong variations in the characteristics of a document will indicate points where the author *intends to* serve a different goal. We use $\mathcal{I}$ to denote the set of all possible intentions and a function $int{:}\mathcal{U}{\to}\mathcal{I}$ that associates every segment to its intention in $\mathcal{I}$. We refer to the text characteristics as *features*, and we will use the term *feature vector* to refer to the values of these features for a segment $s$. Since there is such a close correlation between the features and the intention, given that the intention is only in the mind of the author, it is natural to define it through text characteristics.

*Definition 1:* Given a set $F$ of $n$ features of interest, an *intention* is identified by a feature vector, i.e., a vector of $n$ values, one for every feature of $F$.

The idea of using the features to identify intentions is similar to the idea of using terms to identify topics. In the topic detection literature, the topics of the documents may not be explicitly stated but the terms used in the document are an indication of the topic, and based on this observation, a topic has been defined as a vector of terms [5].

**Problem Statement.** Given a collection $\mathcal{D}$ of documents, and a reference document $d_q$, find those $k$ documents in the collection that are most likely to be related to the reference document $d_q$, i.e., those documents that will most likely be of interest to a user that already considers $d_q$ being of interest. The specific task is referred to as *document matching*.

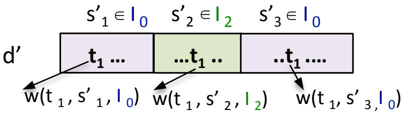

Fig. 1. Weighting for the same term in different intention clusters.

## III. Summary of Intention-based Matching

All the details are found in the full version of the paper [4].

**Segmentation of posts**. The *first challenging task* of Intention-based Matching is to identify the segments of the forum posts. We resort to text features (characteristics), and use the term *communication means* (CM for short) to refer to groups of such features whose variation can identify a passage from one segment to another. We made this choice after realizing that the style, tone, brevity, verb tense and other grammatical characteristics may serve as indicators of a change in the message that the author is trying to communicate. The idea of using communication means for capturing the intention of a segment (or intended message) is analogous to the idea of using keywords to represent a topic. Similar to the way that a variation in a weighted vector of words signals a change in the topic [6], [7], a variation in a vector of text features signals a change in the intended message. By exploiting the communication means, the system identifies the different segments within each forum post and splits the forum post into these segments by employing our *border selection mechanisms* and *coherence/depth functions*.

**Segment grouping**. The *next step* is to recognize segments that are intended for the same goal (or purpose). We actually need to create groups such that segments with similar intentions end up in the same group and segments with different intentions in different groups. Since the actual intention is not known but we have modeled it through a vector of features, a natural choice for creating the desired groups is to perform clustering on the feature vectors corresponding to the intentions of the segments. We have investigated how Communication Means can be converted to feature vectors that capture the different intentions.

**Matching**. Having formed the intention clusters of a document collection $\mathcal{D}$, we can retrieve the top related documents in the collection given any reference document $d_q$ . To measure the *relatedness* of the reference document $d_q$ to another document $d'$ *with respect to a specific intention* $I$, it is enough to measure the relatedness of the respective segment $s'$ of $d'$ in the cluster $I$, to the respective segment $s_q$ of $d_q$ in that same cluster. For computing this relatedness any text comparison can be employed. We devise a version that is somewhere between the original TF/IDF and the BM25. The used weighting scheme is adjusted to consider the intention of the segment where the term is found. Fig. 1 illustrates our weighting approach. In the document $d'$, with $S^{d'}$=$\{s'_1$,$s'_2$,$s'_3\}$, term $t_1$ is weighted differently when found in segment $s'_1$ than when found in segment $s'_2$. Since $s'_1$ is assigned to intention $I_0$, term $t_1$ is weighted based on the terms of $s'_1$ and the respective term index that has been built considering all the segments in $I_0$. On the other hand, the weight of $t_1$ in segment $s'_2$ is estimated considering $s'_2$ and the index built on the segments of $I_1$. The segments with the *highest individual scores* are then selected and their scores are combined to compute a score that indicates how the forum post at hand is believed to be related to other existing forum posts, and based on this score we select the top-k posts (Matching considering all intentions).

## IV. Evaluation & Results

We have evaluated all the steps of our method on the recommendation of related posts i.e., segmentation, identification of segments with the same intention and comparison of the posts based on similarity across segments of the same intention. We used real datasets of posts from forums in three different domains: 111K posts from a product *support* forum (HP Forum), 32K posts of hotel reviews from a *travel* forum (TripAdvisor) and a dump of a well-known computer *programming* forum (StackOverFlow) consisting of  1.5M.

We first needed to see whether the segmentation task we perform makes sense. We have verified the existence of segments in forum posts despite their relative short size and informal writing style, posts *can be naturally divided into segments, with each conveying a different intention* through a user study where users were asked to segment and label forum posts. Regarding the segmentation step, we have also experimentally contrasted alternative features, border selection mechanisms and coherence/depth functions. In addition, we have evaluated our overall approach comparing its effectiveness, in terms of precision, to two baseline methods that are not using any segmentation (a variation of BM25, and a matching based on LDA topics with Gibbs sampling) and two alternatives of our approach where the used segmentation and grouping methods are other than the intention-based ones (two different variations: Content-MR, SentIntent-MR). Our Intention-based matching method, (*IntentIntent-MR*), retrieves the most lists with the largest number of related posts in the first two datasets. Moreover, for the StackOverFlow dataset, it reduces the lists with no true positives (mean precision 0) by 28.6%. Last but not least, experiments on data of different sizes showed that all steps scale well and matching can be performed online.

## 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. 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] 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.

[4] D. Papadimitriou, G. Koutrika, Y. Velegrakis, and J. Mylopoulos, "Finding related forum posts through content similarity over intention-based segmentation," *IEEE TKDE.*, vol. 29, no. 9, pp. 1860–1873, 2017.

[5] I. Hulpus, C. Hayes, M. Karnstedt, and D. Greene, "Unsupervised graph-based topic labelling using dbpedia," in *WSDM*, 2013, pp. 465–474.

[6] M. Hearst, "Texttiling: Segmenting text into multi-paragraph subtopic passages," *Computational Ling.*, vol. 23, pp. 33–64, '97.

[7] H. Misra, F. Yvon, J. M. Jose, and O. Cappe, "Text segmentation via topic modeling: an analytical study," in *CIKM*, 2009, pp. 1553–1556.