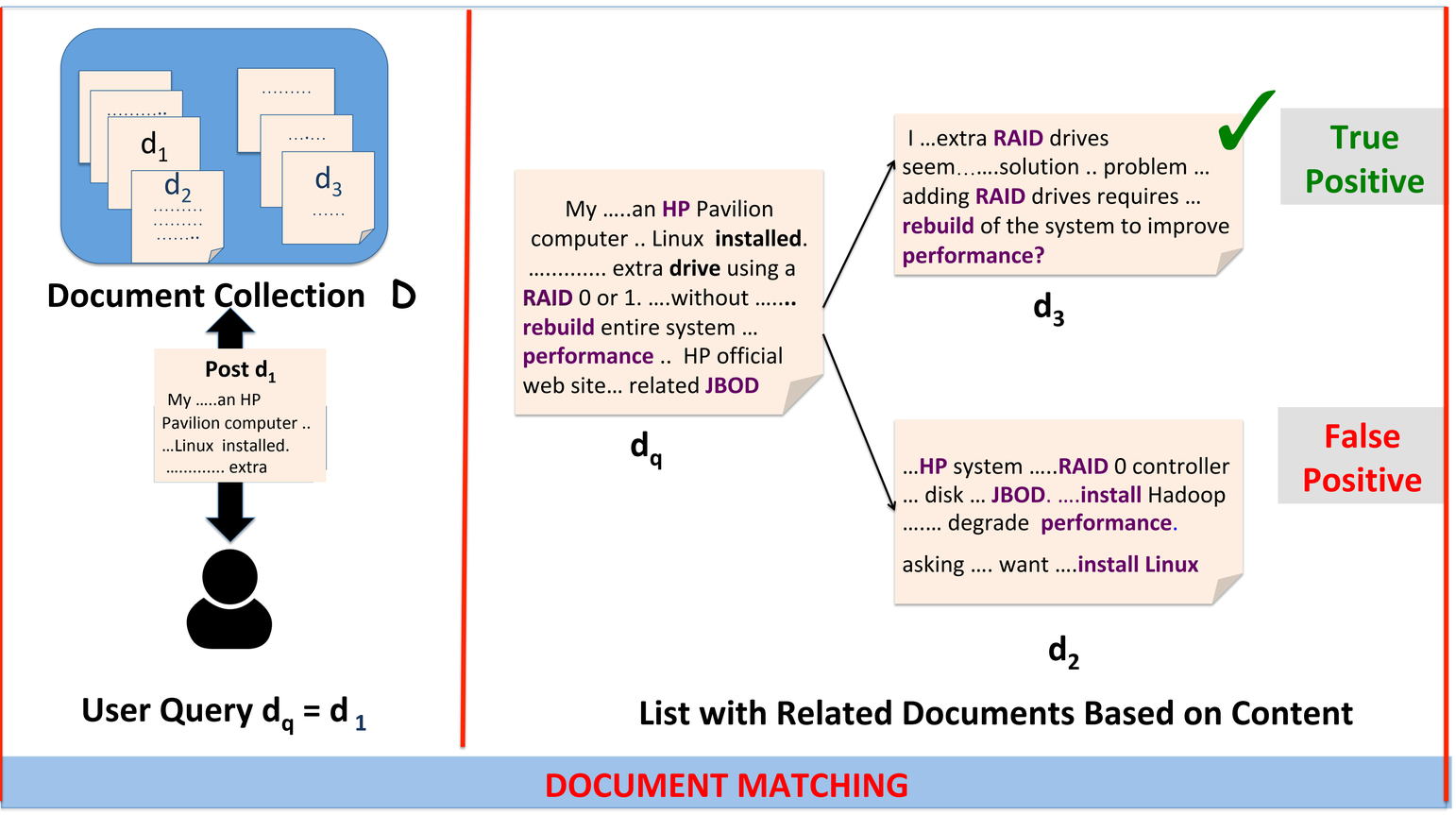
ICDE 2018 | TKDE 2017

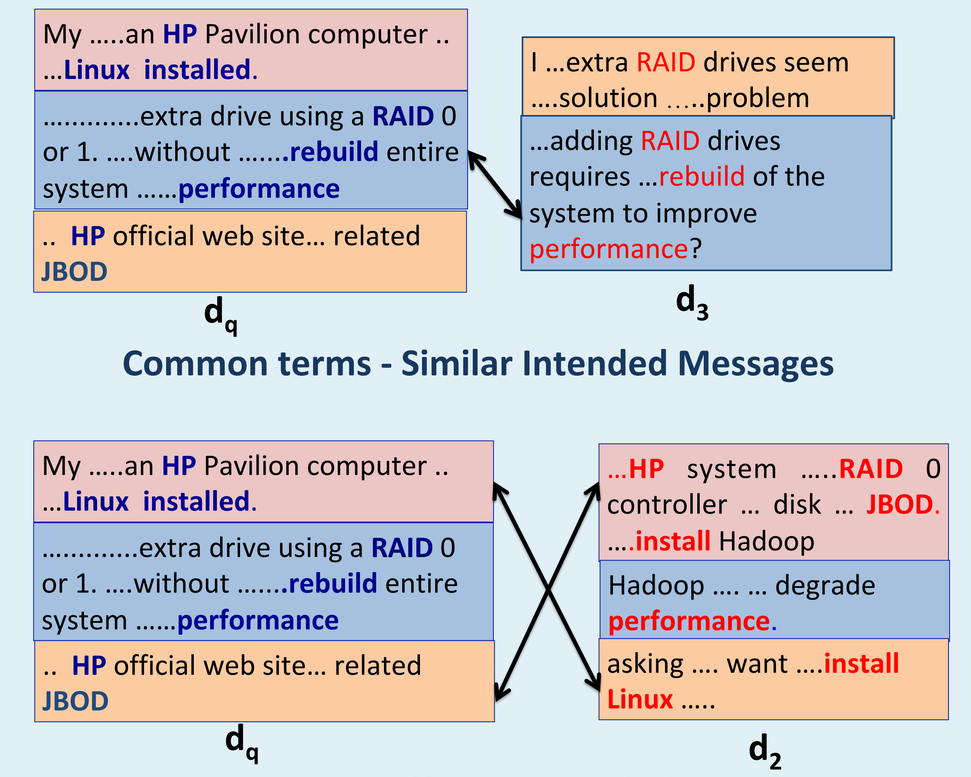
Info: https://dipapadimitriou.github.io/research.html

Finding Related Forum Posts through Content Similarity over Intention-Based Segmentation - Extended Abstract

Dimitra Papadimitriou(*), Georgia Koutrika(**), Yannis Velegrakis(*), and John Mylopoulos(***)

(*) University of Trento (**) Athena Research Center (***) University of Ottawa





The goal for which a piece of text is intended

- may not be explicitly stated
- is reflected into the characteristics of the text

Intention Definition. 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 .

Similarly to the idea of using terms to identify topics, features identify intentions.

INTENTIONS

Content similarity can more accurately determine relatedness, if focused on parts of the forum posts **intended to serve a similar goal.**

- SEGMENTATION OF POSTS
  - Feature Selection
  - Border Selection
- SEGMENT GROUPING
- MATCHING
  - Single Intention Matching
  - All Intentions Matching

INTENTION-BASED MATCHING

Indicators of a change in the goal that the author has written the text for

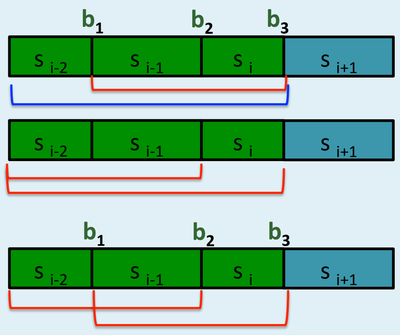
- - Terms, topics
- + Style, tense and other grammatical features

Grouping of features into Categorical variables (Communication Means)

CM_{tense} (Tense)	{present, past, future}
CM_{subj} (Subject)	{I/we, you, it/they/(s)he}
CM_{qneg} (Style)	{interrog., negative, affirmative}
CM_{pasact} (Status)	{passive, active }
CM_{pos} (Part of Speech)	{verb, noun, adj/adverb}

Intention-based Segmentation is based on the distribution of features not the features per se.

FEATURE SELECTION



For a Communication Means CM_r

- **D[j]:** Number of Occurrences of j^{th} value of CM_r (where D is the Distribution Vector for CM_r)
- **All:** Total number of occurrences of any value of CM_r
- **|D|:** Number of value domain of CM_r

$$div_{CM_r}(s_i) = -\sum_{j=1}^{|D|} \frac{D[j]}{All} * log(\frac{D[j]}{All}) \quad coh(s_i) = \frac{1}{|CM|} \sum_{r=1}^{|CM|} 1.0 - div_{CM_r}(s_i)$$

$$depth(b_i) = \frac{|coh(s_i) - coh(s)| + |coh(s_{i+1}) - coh(s)|}{2 * coh(s)}$$

$$score(b_i) = (coh(s_i) + coh(s_{i+1}) + depth(b_i))/3$$

Coherence/Depth functions

BORDER SELECTION

Clustering based on distribution of features

- within each segment
- within each document

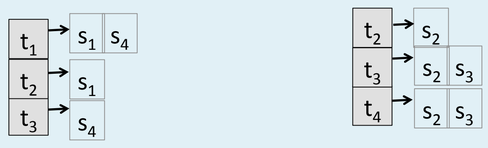
After Segmentation each segment is intended for a goal

Representaion for clustering

Each value of each $CM_{r\ is}$ transformed into two features for the clustering model

I_0		
SEG ID	DOC ID	SEG TEXT
s_1	d_1	t_2 t_1
s_4	d_2	t_2 t_3

I_1		
SEG ID	DOC ID	SEG TEXT
s_2	d_1	t_3 t_2 t_4
s_3	d_2	t_3 t_4



Term Indexing considering Intentions

Matching is performed within each intention cluster, then individual scores are combined to a single score for each document pair.

$$scr(d_q, d', I) = \sum_{\forall t \in s_q} f_{s_q}(t) * w(t, s') * \frac{log(|I| - |I^t|)}{|I^t|}$$

$f_{s_q}(t)$: frequency of the term t in the segment s_q
$|I|$: cardinality of the intention cluster
$|I^t|$: the number of segments in the intention cluster I containing t

SEGMENT GROUPING/FORMING INTENTION CLUSTERS

MATCHING

AUTOMATIC SEGMENTATION VS USER ANNOTATIONS

	HP Forums	Trip Advisor
Offset	Fleiss's κ /Agreement Percentage	
±10 chars	0.20/64%	0.35/71%
±25 chars	0.41/71%	0.44/75%
±40 chars	0.68/77%	0.71/83%

GAIN IN AVERAGE PRECISION

HP Forum	+10%
Trip Advisor	+12%
StackOverFlow	+10.1%

REDUCTION OF LISTS WITH NO RELATED POSTS

HP Forum	-24.5%
StackOverFlow	-28.6%

EVALUATION ON REAL DATASETS

REFERENCES

- D. Papadimitriou, G. Koutrika, Y. Velegrakis, and J. Mylopou- los, "Finding related forum posts through content similarity over intention-based segmentation," *IEEE TKDE.*, vol. 29, no. 9, pp. 1860–1873, 2017.
- J. Jeon, W. B. Croft, and J. H. Lee, "Finding semantically similar questions based on their answers," ser. SIGIR '05. NY, USA: ACM, 2005, pp. 617–618.
- 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.
- I. Hulpus, C. Hayes, M. Karnstedt, and D. Greene, "Unsuper- vised graph-based topic labelling using dbpedia," in *WSDM*, 2013, pp. 465–474.
- M. Hearst, "Texttiling: Segmenting text into multi-paragraph subtopic passages," *Computat. Ling.*, vol. 23, pp. 33–64, '97.
- H. Misra, F. Yvon, J. M. Jose, and O. Cappe, "Text segmentation via topic modeling: an analytical study," in *CIKM*, 2009, pp. 1553–1556.