Temporal Summarization of Knowledge Evolutionary Trend

ABSTRACT

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Social and behavioral sciences

General Terms

Human Factors, Measurement

Keywords

Crowdsourcing, Human computation, Knowledge sharing

1. INTRODUCTION

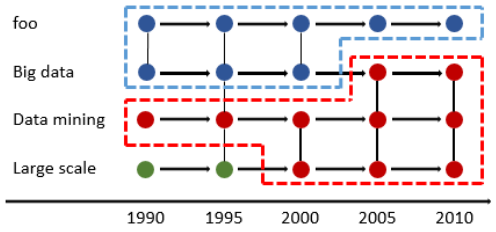
2. METHOD

We choose an optimal summarization of the knowledge evolutionary trend base on *content coverage* and *burstiness* of words, the method yields intuitive coarse-level summarization of the temporal dynamics of a given research topic.

The four-step approach are summarize below:

Step.1 Expert finding We use a community-based summarization as we first find a core community (a group of experts) related the topic, and then aggregate each member's research work as the as the documents to summarize. To avoid some authoritative experts dominate the research area and introduce potential bias, we normalize each member's contribution averagely.

Step.2 Knowledge concept extraction With the retrieved document collection, we extract terms mentioned in each document and construct a network G, where a term w_i appears in a time slide t is associated with a node n_i^t.



There are two types of edges in the graph:

- The concepts in the same time slide are connected with mutual info edge. For an arbitrary edge (n_i^t, n_j^t), the weight from n_i^t to n_j^t is $W(n_i^t, n_j^t) = \frac{I(n_i^t, n_j^t)}{I(n_i^t, n_i^t)}$, where $I(n_i^t, n_j^t)$ indicates the mutual information between w_i and w_j within the time slide t.
- Nodes corresponding to the term w_i within two adjacent time slides t and $t+1$ are connected by an evolving edge (n_i^t, n_i^{t+1}), indicates the knowledge concept is evolved from itself in the last time slide. The weight of the evolving edge is a parameter controlling the temporal smoothness of the summarization.

Step.3 Knowledge concept selection The following task is to choose a set of terms best summarize the whole research area. We use two measures: *Coverage* and *Burstiness* to choose the terms of interest.

- We model burstiness by assuming the arrival of words as a unknown binomial distribution, and use χ^2 tests to check for significant association between word and time periods. We calculate the contingency table as below and $\chi^2 = \frac{(ad-bc)^2 n}{(a+b)(c+d)(a+c)(b+d)}$.

-	W	$\bar{W}$
t	a	b
$< t$	c	d

- We use a influence maximization based model to model information coverage of knowledge concepts. More specifically, we use a linear threshold model, where influence probability from n_i^t to n_j^t is defined as

$$P_{i,j}^t = \alpha P_{i,j}^{t-1} + (1-\alpha) W(n_i^t, n_j^t)$$

, and the activate threshold of n_i^t is

$$\Theta_i^t = \beta \Theta_i^{t-1} + (1-\beta) e^{\sum wX}$$

, where X is a feature vector and w is the weight. The problem is to choose a set of k term that maximize the content coverage.

Step.4 Trend partitioning After choosing a set of summarization words, we further grouping the rest of the words into clusters, and finally we can use the clusters over time to indict evolving trend and generates the highly intuitive visualization.

3. APPLICATION

4. RELATED WORKS

5. CONCLUSION