Mining Temporal Relationships among Categories

Saket S. R. Mengle
Information Retrieval Lab
Department of Computer Science
Illinois Institute of Technology
Chicago, Illinois, USA
saket@ir.iit.edu

Nazli Goharian
Information Retrieval Lab
Department of Computer Science
Georgetown University
Washington DC, USA
nazli@cs.georgetown.edu

ABSTRACT

Temporal text mining deals with discovering temporal patterns in text over a period of time. A Theme Evolution Graph (TEG) is used to visualize when new themes are created and how they evolve with respect to time. TEG, however, does not represent relationships among themes (or categories) that share same timestamp. We focus on identifying such relationships and represent them in Relationship Evolution Graph (REG). We favorably compare passage misclassification and association rule mining with three existing approaches, namely KL divergence (KLD), Consistent bipartite spectral co-partitioning graph (CBSCG) and document misclassification. Our evaluations indicate that association rule mining approach statistically significantly (99% confidence) outperforms the other existing approaches, while passage misclassification approach is the second most effective approach.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**] Clustering

General Terms

Algorithms, Experimentation.

Keywords

Temporal Mining, Text classification, Relationship Evolution

1. INTRODUCTION

Discovering the evolving themes/categories, and the evolving relationships among them are of interest in many applications including in digital libraries. Mei and Zhai [2] proposed a structure called *Theme Evolution Graph* (*TEG*) that represents the theme evolution across different timestamps. The example of *TEG* in Figure 1 represents that the theme Θ_1 evolves into theme Θ_5 in timestamp *T2* and evolves into theme Θ_9 in timestamp *T3*. We are interested to identify relationships among themes or categories that share the same timestamp. We represent such relationships using our proposed structure called *Relationship Evolution Graph* (*REG)*. The example of *REG* in Figure 2 shows that Θ_1 is related to Θ_2 in all three timestamps. Such relationships are called static relationships as they do not change with respect to time. *REG* also shows the dynamic relationships, namely, those that are not constant across all timestamps (pre-defined window of time). Examples of such in Figure 2 are Θ_1 that is related to Θ_5 in two timestamps (*T2* and *T3*) and is related to Θ_3 and Θ_6 in timestamps *T1* and *T2*, respectively. Thus, *REG* represents the span of relationships, that is, the relationship of any given theme or category to the others within each timestamp. *REG* can be applied in various domains such as news filtering, digital libraries and information security, in all of which, discovering the relationships between categories in a given timestamp and its evolution over time is beneficial.

SAC'10, March 22-26, 2010, Sierre, Switzerland.

Figure 1. Example of Theme Evolution Graph (TEG)

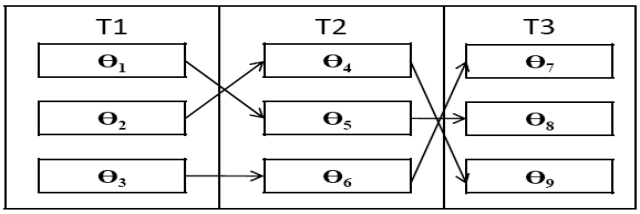
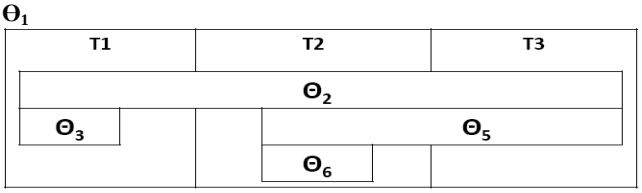
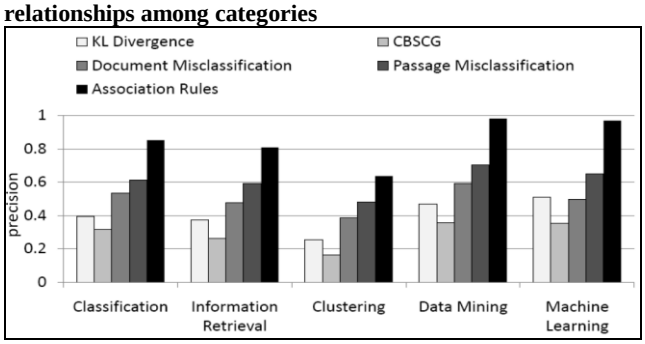


Figure 2. Example of Relationship Evolution (REG) Graph for Θ_1



We propose two approaches to discover relationships among categories, namely passage misclassification and association rule mining and compare them with three existing approaches namely, KL divergence (*KLD*) [2], Consistent bipartite spectral co-partitioning graph (*CBSCG*) [1] and document misclassification [3]. *KLD* between categories c_i and c_j measures the additional information that exists in c_j but is absent from c_i. [2] utilizes *KLD* as a distance measure to discover relationships among themes. *CBSCG* is a clustering based approach that co-partitions a tripartite graph between documents, terms and categories to generate a hierarchy that represents relationships among categories. The *document misclassification* approach assumes that most misclassifications (false positives and false negatives) generated during the process of document classification occur in categories that are closely related to each other. Hence, a relationship is identified between two categories c_i and c_j when the highest number of the false positives or false negatives for category c_i occurs in category c_j.

Figure 3. Comparison of various approaches to identify relationships among categories



2. PROPOSED APPROACHES

2.1 Passage Misclassification

Passage misclassification approach like *document misclassification* approach [3] utilizes the misclassification information, however, it utilizes the passages within documents to classify and generate such information as opposed to using the whole document. Our premise is that although an entire document may not be misclassified, passages within that document may be misclassified into categories that are indeed *related* to the actual category of that document. This additional information leads to an improvement in precision over *document misclassification* approach. Similar to [3], we used a Naïve Bayes classifier for our classification task.

2.2 Association Rule Mining

To improve the detection effectiveness of relationships among the text categories, our approach derives relationships among categories using association rule mining. We calculate the support and confidence between each two categories. The *support* (c_i, c_j) (Formula 2.1) for categories c_i and c_j is defined as the proportion of publications (N) in the dataset that contain both c_i and c_j ($\sigma(c_i \cup c_j)$). Confidence (Formula 2.2) is defined as a probability that category c_j exists when a document belongs to category c_i.

$$Support(c_i \Rightarrow c_j) = \frac{\sigma(c_i \cup c_j)}{N} \quad .. 2.1$$

$$Confidence(c_i \Rightarrow c_j) = \frac{\sigma(c_i \cup c_j)}{\sigma(c_i)} \quad .. 2.2$$

We only identify relationships among categories whose support and confidence is above an empirically determined threshold.

3. EXPERIMENTATION

3.1 Framework

We explore the usefulness of *REG* in digital libraries. Our datasets consist of the last ten years publications from SIGIR, KDD and CIKM conferences. We identify categories from the *keywords* section of publications. We only select the fifty most frequently occurring categories for each timestamp.

To validate our approach, we manually evaluate the correctness of predicted relationships and report precision for *REG*s generated by our approaches. We evaluate five *REG*s for five categories, namely *information retrieval*, *data mining*, *machine learning*, *classification* and *clustering*.

3.2 Results

We observed that among the earlier efforts, document misclassification statistically significantly (99% confidence) outperforms *CBSCG* (Avg. Improvement: 21.63%) and *KLD* (Avg. Improvement: 10.66%) in all the five *REGs* (Figure 3). This stems from the fact that both *KLD* and *CBSCG* are unsupervised approaches, whereas *document misclassification* approach is a supervised approach.

Passage misclassification approach statistically significantly (99% confidence) outperforms *document misclassification* approach (Avg. Improvement: 10.35%) for all five *REGs*. Effectiveness of misclassification approaches are dependent on the number of misclassifications that are generated during the process of text classification. The average number of publications assigned to a category in our dataset is 16.43. Hence, the document misclassification bases its decision on very few misclassifications per category. *Passage misclassification* approach divides a document into passages and classifies each passage. Hence, the number of misclassifications generated using passage classification is much higher than in document classification approach. As both approaches use the same text classification model for predicting categories, the quality of misclassification is similar. Hence, *passage misclassification* approach performs statistically significantly better than *document misclassification* approach.

Association rule mining approach statistically significantly (99% confidence) outperforms all other existing approaches, namely *KLD* (Avg. Improvement: 44.78%), *CBSCG* (Avg. Improvement: 46.27%) and *document misclassification* (Avg. Improvement: 34.13%), and our proposed approach, *passage misclassification* (Avg. Improvement: 23.87%). This also stems from the fact that the average number of publications assigned to a category is low. The correlation between the precision and number of documents assigned to a given category is higher for *document misclassification* (97.62%) and *passage misclassification* (95.27%) approaches than for *association rule mining* approach (68.67%). Hence, the misclassification approaches perform worse than *association rule mining* approach.

In summary, we introduced a structure called *REG* that represents temporal relationships among categories that share the same timestamp. Our results indicate that using *association rule mining* approach statistically significantly outperforms *KLD*, *CBSCG*, *document misclassification* and *passage misclassification* approaches, while *passage misclassification* approach is the second most effective approach.

4. REFERENCES

[1] Gao B., Liu T., Cheng Q., Feng G., Qin T., and Ma W., Hierarchical taxonomy preparation for text categorization using Consistent Bipartite Spectral Graph Co-partitioning. IEEE Trans. on Knowledge and Data Eng. Vol 17 (9), 2005.

[2] Mei Q., Zhai C., Discovering evolutionary theme patterns from text: an exploration of temporal text mining, ACM Int. Conf. on Knowledge Discovery in Data Mining, 2005.

[3] Mengle S., Goharian N., Platt A., Discovering relationships among categories using misclassification information, Proc. of the ACM 23rd Symp. on Applied Computing, 2008.