Performing Impact Analysis using the Relation Topic based Coupling
Malcom Gethers
Advisor: Denys Poshyvanyk
Computer Science Department
The College of William and Mary
Williamsburg, VA 23185
mgethers@cs.wm.edu

ABSTRACT
The proposed research defines a novel technique which applies an advanced Information Retrieval (IR) method to leverage textual information embedded in software artifacts in order to capture coupling between classes. More specifically, here we propose the use of new emerging IR techniques based on topic modeling, namely Relational Topic Model (RTM), to define a coupling metric useful for impact analysis.

Categories and Subject Descriptors
D.2.8 [Software Engineering]: Metrics

General Terms
Measurement.

Keywords
Coupling, Information Retrieval, Topic Models.

1. INTRODUCTION
Coupling metrics capture the degree of interaction and relationships among source code elements in software systems. A large majority of coupling metrics in the literature rely on structural information, which shows relations such as usage relations between source code elements (e.g. class, method, etc.). Those metrics lack the ability to identify conceptual relationships, which, for example, specify relationships encoded by developers in identifiers and comments of source code. We propose a new coupling metric for object-oriented software systems, which uses Relational Topic Model [6], an unsupervised probabilistic topic modeling technique. RTM, a novel Information Retrieval method, identifies latent topics associated with documents (e.g. source code elements). RTM extends on Latent Dirichlet Allocation (LDA) [3], which has been recently applied for extracting, representing and analyzing latent topics from the source code [2, 13-16, 20].

Evaluation of our novel metric consists of computing both established and our RTM-based coupling metrics on a set of open source software systems. We compare the performance of coupling metrics for impact analysis, an important software maintenance task. While other coupling metrics have been successfully used for impact analysis in the literature we conjecture that proposed conceptual coupling metric captures a unique aspect of coupling overlooked by existing metrics as well as provides good accuracy for impact analysis. This paper summarizes our recent findings [10].

2. RELATED WORKS
One of the active research areas in software engineering is coupling measurement, resulting in metrics based on structural static information [5, 7, 12], dynamic information [1], evolutionary data [9, 22], and textual information [18, 19]. Additionally, coupling metrics have been defined to capture coupling metrics for specific types of software applications, such as knowledge-based [11] and aspect-oriented software systems [21].

3. Using Relational Topic models for coupling measurement in software
In order to perform impact analysis we define a novel coupling metric which leverages RTM to determine the degree to which classes are conceptually coupled. RTM models documents (i.e., classes) as mixtures of topics and models links between documents (i.e., couplings) within a collection of documents (i.e., software system). Identification of potential friends in social networks, pinpointing relevant citations of a technical papers, and identifying relevant webpages for a given webpage are a few purposed applications of the model [6].

3.1 Measuring Coupling using Relational Topic Model
RTM provides a comprehensive model for describing documents (i.e., classes are represented as words from identifiers and comments) and the existence of links between documents based on underlying textual information and other knowledge of the document network. In the context of our application, the binary link indicator, which indicates whether a link exists between two documents (i.e., classes), determines coupling amid any pair of classes. That is, if the model identifies a link between two classes in the corpus with a high probability, we consider these classes to be coupled.
4.1 Research Questions and Results

We define a set of research questions (RQ) we will address to provide insight into the usefulness of RTC.

Table I. Precision (P) and recall (R) results for impact analysis in Eclipse using various cut points

<table>
<thead>
<tr>
<th></th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>P</td>
<td>R</td>
<td>P</td>
<td>R</td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>RTC_C+CCBC_m</td>
<td>21</td>
<td>25</td>
<td>17</td>
<td>35</td>
<td>15</td>
<td>40</td>
<td>13</td>
<td>45</td>
</tr>
<tr>
<td>CCBC_m</td>
<td>19</td>
<td>23</td>
<td>15</td>
<td>31</td>
<td>13</td>
<td>36</td>
<td>12</td>
<td>41</td>
</tr>
<tr>
<td>Absolute gain</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Relative gain</td>
<td>11</td>
<td>9</td>
<td>13</td>
<td>13</td>
<td>15</td>
<td>11</td>
<td>8</td>
<td>10</td>
</tr>
</tbody>
</table>

A term-by-document co-occurrence matrix representation of a software system must be derived in order to use RTM to capture coupling. Subsequently the system is modeled as distributions of topics within the corpus [6]. After obtaining the model, given two classes \( C_1 \) and \( C_2 \) the degree of coupling (i.e., pair-wise RTC\(_C\)) between these is defined as follows:

\[
RTC\_C = RTM(C_1, C_2)
\]

where the function \( RTM() \) returns the probability that a link exists between the classes \( C_1 \) and \( C_2 \). System level coupling or the degree to which a class is coupled to the other classes in the system (i.e., system-level RTC\(_S\)) is defined as:

\[
RTC\_S(C_i) = \frac{\sum_{m=1}^{n} RTM(C_i, C_m)}{n}
\]

where \( n \) is the number of classes in the system.

4. CASE STUDY

In this section we present the design of an empirical case study aimed at comparing RTC with other structural and conceptual coupling metrics and analyzing whether the combination of RTC with existing conceptual coupling metrics improves the accuracy during impact analysis. In total 11 C/C++ and two Java software systems are used in the study (see Gethers and Poshyvanyk for details [10]).

Nine existing structural metrics are considered when evaluating our novel metric (CBO, RFC, MPC, DAC, ICP, ACAIC, OCAIC, ACMIC, and OCMIC). Two system-level conceptual coupling metrics, namely CoCC and CoCC\(_m\) (and their pair-wise versions, CCBC and CCBC\(_m\)) [18], are also considered in this study. Structural coupling measures used in this study were computed using \textit{Columbus} [8] while CoCC and CoCC\(_m\) measures were obtained using the \textit{IRCDM} tool [18]. RTC on the other hand was computed using the \textit{lda}\(^1\) package available for the \textit{R-project}\(^2\).

4.1.1 RQ\(_1\) – Results of principal component analysis of the metrics data

Our first research question focuses on using Principal Component Analysis (PCA) to determine if our proposed metric captures a dimension in the data unexplained by previous coupling metrics. Our use of PCA adheres to the methodology used in our prior work [29, 31, 36]. Using a total of 10 open source software systems, metrics are computed for 978 classes, revealing six orthogonal dimensions (90.97% of the variance in the data set) which explain the 12 coupling measures. We observe that one of the six dimensions is highly correlated (0.93) with only RTC, indicating our metric captures a unique dimension in the data. It is our conjecture that this property is present because of the different underlying mechanism used to analyze the conceptual information (i.e., RTC uses RTM as opposed to LSI which is the case for CoCC).

4.1.2 RQ\(_2\) – Comparing results of RTC with structural and conceptual coupling metrics for impact analysis

Answering RQ\(_1\) allowed us to show that RTC captures a unique dimension in the data revealing the fact that RTC explains an aspect of the data overlooked by existing metrics. Insight into the practical usefulness of this metric is required to answer RQ\(_2\), therefore we apply RTC to the task of impact analysis in source code [4]. Results of applying the various metrics to impact analysis for \textit{Mozilla} indicate that that RTC\(_C\) and CCBC\(_m\) are the best coupling metrics, as they significantly outperform the structural metrics. Results for RTC\(_C\) reveal lower precision when compared to CCBC\(_m\) at lower cut points. Note that RTC\(_C\) yields better recall for cut points 50 and greater. In answering RQ\(_1\) we show that RTC captures a unique dimension, and RQ\(_2\) shows RTC is more accurate for impact analysis than the other metrics. Therefore we conclude that it coupling measurement research benefits from the introduction of RTC. Based on the findings of this work, we explore if combining RTC\(_C\) with CCBC\(_m\) (other metric which yielded good performance) results in any additional improvements in accuracy of coupling-based techniques for the task of impact analysis.

4.1.3 RQ\(_3\) – Results of combining RTC with another conceptual coupling metric for impact analysis

To answer RQ\(_3\) we performed impact analysis on Eclipse, using the combinations of RTC\(_C\) and CCBC\(_m\) (using affine transformation with equal weights to both techniques as in prior work [17]. Table I shows that coupling metrics, RTC\(_C\) and CCBC\(_m\), provide comparable accuracy, as shown in the results related to RQ\(_2\). Although, in this section results indicate that the

\(^1\) http://cran.r-project.org/web/packages/lda/
\(^2\) http://www.r-project.org/
combination of the metrics results in accuracy superior to either standalone technique. We confirm the statistical significance of such results using Wilcoxon's signed-rank test (details appear in prior work [10]). Subsequent to answering RQ3 we are able to conclude that RTC represents a novel coupling metric useful and practice (i.e., predicting classes which co-change).

5. CONCLUSION

The paper defines a novel measure for the relational topic based coupling of classes, which is theoretically and empirically validated. An extensive case study reveals the usefulness of RTC in practice and its unique theoretical properties.

6. ACKNOWLEDGEMENTS

This research is supported by the Virginia Space Grant Consortium.

7. REFERENCES


