Analyzing the Effectiveness of Pruning and Grouping Methods Used in Literature-Based Discovery Tools

Fatih Mehmet GULEC, Tahir BICAKCI, Ebru AKCAPINAR SEZER, Hayri SEVER
Department of Computer Engineering
Hacettepe University
Ankara, Türkiye
fatihmg@hacettepe.edu.tr

Vijay V. RAGHAVAN
The Center for Advanced Computer Studies
University of Louisiana
Lafayette, LA
raghavan@louisiana.edu

Abstract—LBD tools enable the establishment of relationships between concepts appearing in scientific articles in the biomedical field and the generation of new hypotheses via the examination of these existing relationships. In this paper, we study the effectiveness of generally accepted grouping and eliminating logics used in LBD tools. This work is performed in the context of Lit2Info, a system that we have developed as an LBD tool. Significant performance variations over that of popular pruning and grouping methods are reported. One of our findings is that there is no positive or negative impact of pruning of closely related terms in links, which is one of the pruning methods commonly applied to decrease the number of irrelevant hypotheses. More importantly, we find that the pruning methods used in this connection actually lead to a decrease in the effectiveness of the hypotheses generation process.

Keywords—Text Mining, Knowledge Based Discovery, Literature Based Discovery

I. INTRODUCTION

In the light of information obtained from fact databases and fresh publications, ideas to make research on current and specific issues emerge [2]. A researcher generates specific ideas heuristically, as the result of brainstorming with and matching of concepts that arise from the information obtained [3, 4]. The fact that around 2000 new articles are published in PubMed every day [1] demonstrates the reality that it is not possible for the researchers to keep up with contents in all publications that are directly or indirectly related to their expertise and discern emerging research trends. Consequently, it is highly desirable to design approaches that model research idea generation strategies by simulations in computer systems.

Literature-Based Discovery (LBD) targets at producing “new” information by bringing together various studies in a complementary manner based on the previously published information [7]. LBD, in the context of the biomedical field, seeks to establish relations between concepts over scientific articles and produce novel and accurate relationships based on the examination of existing relationships [6]. Whereas novelty ensures that the hypothesis asserted is not included in the previous works, accuracy demands that derived connections between terms are significant and accurate.

In the experiments we carried out with Lit2Info, which is a system we have developed as a LBD tool, we examined the pruning methods asserted to prevent the presentation of irrelevant hypothesis in LBD tools. The fact that the pruning methods target the elimination of insignificant hypotheses suggests that our contributions in this paper address issues that directly affect the usability of the tools. Whereas the pruning methods asserted for this purpose have a great significance, the fact that the efficiencies are not measured previously restricts the study of accuracy only at the logical level. In the experiments we carried out on Lit2Info, the experimental analysis of interaction between result accuracy and the pruning methods is done for the first time.

II. RELATED WORKS

The 1986 study of Swanson constitutes the first example in LBD field [9, 10]. In this work, Swanson has examined relationships derived from publications prepared in disjoint areas, crossed those relationships and derived new possible research hypotheses. As a concrete example, while Swanson was working on existing research papers on Raynaud disease, he noticed that, “blood viscosity and platelet clustering is high” is a typical characteristic of Raynaud patients. On the other hand, he realized that “the fish oil and its active ingredient eicosapentaenoic acid, decreases clustering of blood platelets and blood viscosity”. Using the transitivity property between these two relationships, Swanson asserted that, as a new hypothesis, fish oil could potentially be used in the treatment of Raynaud disease.

Lindsay and Gordon applied the term frequency and inverse document frequency methods to this field, within the context of information retrieval [12]. Xie et al. brought a different perspective to the methods of producing new hypothesis from the existing hypothesis, and added the substitution and chaining rules to the ordinary transitivity-based hypothesis generation [13].

Pratt et al. reported that the strategy of grouping terms that have similarity in terms of meaning, is effective for the elimination of unnecessary or irrelevant hypotheses [14]. According to term grouping approach, terms that are combined on the basis of meaning are treated as a group. As an example, terms, such as “serum magnesium” and “magnesium level,” were grouped and processed as a single term. In a study published in 2005, the more general expressions of the terms are determined and eliminated using the ancestor relationship in MeSH term tree [15]. While ranking the terms, emphasis was put on terms that are determined to be used relatively more frequently in relevant documents, and they were presented
to the end user at higher ranks [16]. In a 2009 study, methods towards comparison of results of studies carried out in the field of literature based information discovery were recommended [17].

III. OVERVIEW OF Lit2INFO AND EXPERIMENTAL DESIGN

The input of LBD systems comprises of a start term, for which new relationships are to be determined, and the semantic constraints to be used for preferred relationship derivation. The output of system is the list of the target terms that are indirectly related to the start term based on relationships explicitly stated in the existing publications.

The workflow of the system starts by supplying the articles including the starting term, which is the input parameter and has been shown as Term A in “Fig. 1”. MeSH terms are derived from the articles and linking terms (B terms), which are co-occurred with the starting term in these articles, are determined. These relationships are demonstrated as lines drawn between A terms and B terms in “Fig. 1”. In the second phase, new articles are obtained by queries built using linking terms and the target terms are obtained from these articles. Target terms are represented as term C in the “Fig. 1”. In order to make sure that the hypothesis to be generated by the system is new, it is required to eliminate the terms among the target terms that are directly related to the starting term. For that purpose, terms that are found at the intersection of target terms and linking terms are removed from the target term set.

At the last step of workflow, hypotheses (relationships between start and target terms), which are the outputs of the system, are listed according to their relevancy. Relevancy is given as a function of (1) number of linking terms involved along the derivation paths, and (2) the average of frequency of linking terms used to build that hypothesis.

While designing the Lit2Info, mentioned phases are implemented as modules. There are modules for retrieving the articles, pruning and grouping the terms, establishing relationship between terms and listing of results. The interaction between modules is given in “Fig. 2” and the flow diagram of system is given in “Fig. 3”.

A. Pruning

Pruning processes are carried out in order to ensure those irrelevant hypotheses are eliminated as much as possible. Pruning procedures are performed with three different methods:

1) Terms that are considered to be highly general are removed from the linking and target term lists. To ensure that, terms used in more than 10,000 article are pruned. These terms are marked as general term and used to build the stop word list.

2) Terms that have close relationship with the starting term are removed from the list. For example, when migraine headache is the starting term, the retinal migraine term should not be used as linking term. Being closely related is determined by using the ancestors and first offspring of a term given in MeSH term tree.

3) Establishing relationship between terms, without regard to the location of the terms in the documents and their meanings, will result in the addition of spurious relationships, which should be eliminated in subsequent steps [17]. In order to avoid this situation, relationships whose semantic types do not conflict are included in the result. It is up to the user to determine which types of relationships can be established and which of these could be eliminated, using UMLS ontology.

Figure 1. Matching of Concepts (cited from Weeber. [11])

Figure 2. Modules of the applied LBD – Lit2Info

Figure 3. Work flow diagram of the system
B. Grouping

Identifying the significance degrees of terms separately—regardless of whether the similar terms are used for the same purpose—will lead to irrelevant results. For this purpose, those that are similar among the terms will be grouped and handled as a single term. For example, the term “magnesium” could not obtain sufficient point among the documents obtained with the term “epilepsy”. The terms “serum magnesium”, “magnesium level” and “magnesium deficiency”, which are closely related to the term “magnesium” will not collect sufficient points on their own. However, ranking could be done more effectively if these terms are grouped and their degrees of importance are combined.

C. Method for Testing and Evaluating the System Outputs

It was seen that there was no generally applicable method as in the case of other disciplines and information retrieval systems for the analyzing and testing the accuracy and validity of results produced by LBD tools [18]. This situation arises from the lack of a quality test database (Benchmark Data Set) specialized for LBD works. It is an inevitable requirement to compare the results of an algorithm or system with others, to test the efficiencies, accuracy and completeness of the results. The test method frequently encountered in LBD studies is framed around the hypothesis generation method used by Swanson. The evaluations have focused on re-finding the relationship between Raynaud diseases and fish oil using documents prior to the year in which Swanson has originally found this. Criterias of methods to analyse the results of LBD systems are mentioned in [17].

According to a different approach, which is based on the idea of using history as a guide to future, articles were divided into two chronological parts; articles that are published before a certain date are considered to be pre-cutting, and the articles that are published thereafter are called after-cutting [17]. In this method, target terms are provided from pre-cutting articles. The experimental question is whether the target terms obtained from pre-cutting articles have a direct relationship with the start term in the after-cutting articles (the situation of being the linking term in terms of literature based information discovery). This assessment approach is applied in Lit2Info and explained in the work logic specified in “Fig. 4”.

In the comparison of clusters, the precision and recall concepts, which are measures used in evaluating information retrieval systems method, are used (“Eq. (1)” and “Eq. (2))

\[
\text{Precision} = \frac{|T_b \cap G|}{|T_b|} \\
\text{Recall} = \frac{|T_b \cap G|}{|G|}
\]

IV. EXPERIMENTS AND RESULTS

In our study, the efficiency of grouping and pruning methods are evaluated. By calculating the precision and recall values, results are compared. The comparisons are based on 100 randomly selected terms. Precision and recall values of different experiment groups are interpreted according to paired samples t-test. Acceptance / rejection status of hypothesis is determined taking into account the confidence bounds of 0.01 (t=2.626).

A. Grouping experiment

Initially, the grouping method has been tested in a series of experiments. Grouping on the linking term directly affects the ranking of target terms. However, due to the fact that linking term grouping has no effect on the determination of target terms, grouping experiment was not made on the linking term. In order to examine the effects of presentation of target terms by grouping, precision and recall values of target ungrouped outputs and grouped outputs are calculated separately. Within each calculation gold standards have been redefined according to being grouped or not.

<table>
<thead>
<tr>
<th>TABLE I.</th>
<th>EVALUATION OF GROUPING METHODS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
</tr>
<tr>
<td>Target terms grouped</td>
<td>0.0446</td>
</tr>
<tr>
<td>Target terms ungrouped</td>
<td>0.0585</td>
</tr>
<tr>
<td>T value</td>
<td>4.595</td>
</tr>
</tbody>
</table>
When the precision and recall measurements are compared for two groups, the hypothesis of “there is no difference between groups” is rejected in the confidence bounds of 0.01. According to this, with target term grouping, both the precision and recall values are negatively affected.

**B. Pruning experiment**

Pruning processes are performed on linking terms in order to eliminate the irrelevant hypothesis in LBD tools. It is considered that improving the quality of the pruning process is the most important factor that impacts the usability of LBD tools. Pruning that remain below the optimum level will cause production of numerous meaningless hypotheses leading to such over-burdening that is difficult for the user to cope with. In other words, this will lead to a decrease in accuracy. In case that pruning is made at a higher level than the optimum level, inference of novel relations will be conceded.

From the point of view of ensuring accuracy and novelty balance, it becomes important to examine pruning methods separately and measure their performance impact. For that purpose, results obtained by, applying and not applying, 3 different pruning methods are provided for randomly selected 100 terms (general term pruning – GTP, closely-related term pruning – CRP and semantic type confliction pruning - SP) are compared.

**TABLE II. EVALUATION OF THE PRUNING METHODS**

<table>
<thead>
<tr>
<th>Measurement Groups</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTP</td>
<td>0.0239</td>
<td>0.7309</td>
</tr>
<tr>
<td>GTP + SP</td>
<td>0.0069</td>
<td>0.1669</td>
</tr>
<tr>
<td>GTP + CRP</td>
<td>0.0236</td>
<td>0.7307</td>
</tr>
<tr>
<td>GTP + SP + CRP</td>
<td>0.0068</td>
<td>0.167</td>
</tr>
</tbody>
</table>

For the comparison of measurements, the 4 experiment groups are subjected to paired samples t-test. GTP + SP and GTP+CRP comparison is not included since it does not deal with the situation of existence and absence of one of the three methods.

**TABLE III. SIGNIFICANCE TESTS FOR COMPARISON OF PRUNING METHODS**

<table>
<thead>
<tr>
<th>Group1</th>
<th>Group2</th>
<th>M</th>
<th>t</th>
<th>Hyp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTP</td>
<td>GTP+SP</td>
<td>P</td>
<td>8.364</td>
<td>rejected</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>21.442</td>
<td></td>
<td>rejected</td>
</tr>
<tr>
<td>GTP</td>
<td>GTP+CRP</td>
<td>P</td>
<td>1.136</td>
<td>accepted</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>0.332</td>
<td></td>
<td>accepted</td>
</tr>
<tr>
<td>GTP</td>
<td>GTP+SP+CRP</td>
<td>P</td>
<td>8.438</td>
<td>rejected</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>21.430</td>
<td></td>
<td>rejected</td>
</tr>
<tr>
<td>GTP+SP</td>
<td>GTP+SP+CRP</td>
<td>P</td>
<td>1.00</td>
<td>accepted</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>0.332</td>
<td></td>
<td>accepted</td>
</tr>
<tr>
<td>GTP+CRP</td>
<td>GTP+SP+CRP</td>
<td>P</td>
<td>8.298</td>
<td>rejected</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>21.514</td>
<td></td>
<td>rejected</td>
</tr>
</tbody>
</table>

Hypothesis Rejected - There is difference between groups. Hypothesis Accepted - There is no difference between groups. M = Measure. P = Precision, R = Recall

Precision and recall values are surprisingly low at first glance. Test method we used relies upon new hypothesis will be found after pre-cut date. But as excepted, all the possible relations are not still established. So, the precision and recall values are insufficient to say the overall performance of the system, but only gives a comparison facility. This fact also shows itself in other studies. For example, Wanda gives precision as 0.08 when she worked on "schizophrenia", and the best recall value is 0.48 for "migraine" in her study [16]. Like as Wanda, Xie and colleagues give precision in the range of 0.01 and 0.25 [13].

**V. DISCUSSION**

In case that grouping is made on the target term, the detail level of hypothesis to be produced as output will be decreased automatically. Both the precision and recall value has decreased significantly with the grouping of results to be determined as new hypothesis. It was observed that this approach led to negative results.

It is observed that the Closely Related Term Pruning (CRP) approach does not have a positive or a negative impact. The fact that there is no difference between groups in comparisons between GTP and GTP+CRP and between GTP+SP and GTP+SP+CRP demonstrates that the CRP approach has no impact.

The most important result is the observation that the semantic pruning does have a negative impact. There occurs a significant fall in both precision and recall values when a semantic pruning is performed. Semantic pruning causes the elimination of linking terms in the production of new hypotheses when these are in opposition with the starting term or target term. The fact that this approach leads to negative results demonstrates that new relationships have emerged due to factors that could not be estimated at all.

**REFERENCES**


