Networked Hierarchies for Web Directories

Nazli Goharian
Computer Science Department
Georgetown University
goharian@georgetown.edu

Saket S.R. Mengle
Dataxu Inc
Boston, USA
smengle@dataxu.com

ABSTRACT

The hierarchical nature of existing Web directories, ontologies, and folksonomies, are known to provide meaningful information that guide users and applications. We hypothesize that such hierarchical structures provide richer information if they are further enriched by incorporating additional links besides parents, and siblings, namely, between non-sibling nodes. We call such structure a *networked hierarchy*. Our empirical results indicate that such a networked hierarchy introduces interesting links between nodes (non-sibling) that otherwise in a hierarchical structure are not evident.

Categories and Subject Descriptors

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

General Terms

Algorithms, Performance, Design, Experimentation

Keywords

Taxonomy, Text classification, Hidden Information

1. INTRODUCTION

Web directories such as Yahoo and Open Directory Project (ODP) classify web pages into document hierarchies. Such hierarchies are useful for effective information management. We are interested to automatically identify non-sibling relationships and generate links among categories that do not share the same parents. An effort to utilize non-sibling relationships was described in [1] and created a *relationship-net*. Although such network provided some information that otherwise was not evident in a hierarchical structure, it was disadvantaged by the fact that the hierarchical characteristics of parent, child, and sibling relationships were, in part, lost. A *networked hierarchy* is a hierarchy that not only maintains the characteristics of a hierarchy, i.e., parent, child, sibling, but also provides links between those non-sibling categories (nodes) that are, indeed to a degree, relevant. A weight is calculated for each such link to indicate the strength of such relationship. Figure 1 shows an example of a *networked hierarchy*. This is a subset of the ODP hierarchy, with the additional links that our system identified between non-sibling nodes. As shown, there is a relationship between non-sibling nodes <Football, Games> and <Graphics, Games> that otherwise was not evident in a category hierarchy.

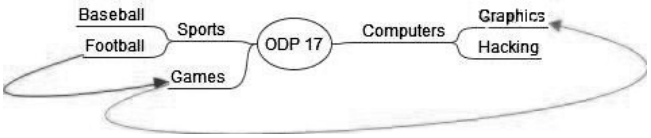

Figure 1. Example of a *networked hierarchy*

2. CONSTRUCTING NETWORKED HIERARCHY

We capitalize on earlier efforts in identifying relationships among categories. Some provide a higher precision and lower recall and some a higher recall and lower precision. Independent of the approach, we validated our hypothesis that such a networked hierarchy is more than a hierarchy. That is, it provides additional information that can be of interest. Next, briefly we explain the approaches used in [1] and [2], namely, using *misclassification information* and *association rule mining*, respectively.

A *Misclassification based approach* [1] is a text classification-based approach that utilizes misclassification information, i.e., false positives and false negatives. These classifications are generated during the process of text classification to detect relationships among categories. A relationship is predicted between category 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 or vice versa. The premise of this approach relies on the finding that categories that mostly are misclassified as each other indeed are relevant. The authors in [1] evaluated the approach favorably. A 5-step approach constructs a confusion matrix for the actual and predicted categories, and then after normalizing the values, identifies false positives (*FP*) and false negatives (*FN*) for each category. Finally, it establishes a relationship between two categories based on their highest false positive and false negative values. Unlike in [1], our goal is to maintain the hierarchy structure of the data while establishing additional relationships and links. Hence, we disregard the misclassifications that occur among siblings, and parent-child categories, and consider only the non-sibling misclassification information in calculating the *FP* and *FN*. A weight is calculated based on the normalized *FP* or *FN* values, depending on which one is the highest, and is assigned to each link (non-sibling relationship).

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

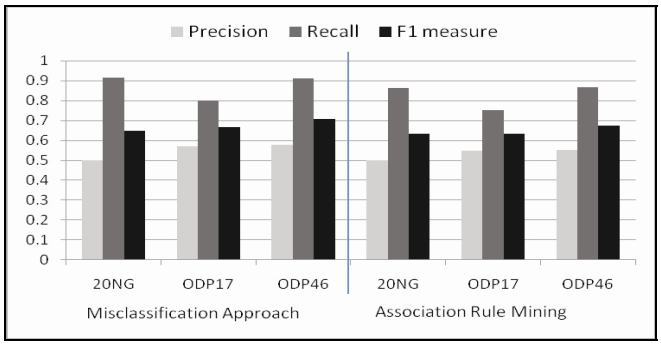

Figure 2. Evaluation of new relationships identified between non-sibling nodes in a hierarchy using misclassification-based and association rule mining approaches

An *Association Rule Mining* based approach calculates the support and confidence between each two categories in the hierarchy to determine additional relationships between candidate non-sibling nodes [2]. A set of categories, $C=\{c_1,c_2..c_n\}$, is considered as an *itemset* during the process of *association rule mining;* the database D is a set of misclassifications, $D=\{c_{actual}, c_{predicted}\}$, with each misclassification having two elements from the *itemset*. The *support* (c_i, c_j) for categories c_i and c_j is defined as the ratio of data that contain both c_i and c_j $(_{\sigma (c_i \cup c_j)})$.

Confidence of a rule with two categories is calculated as the probability of category c_j when a document belongs to category c_i $(c_i => c_j)$ or vice versa $(c_j => c_i)$.

3. RESULTS

We empirically evaluated the effectiveness of such a *networked hierarchy*, using available benchmark datasets that present data in a hierarchy. For that, we used two versions of Open Directories Project namely, ODP17 (8,500 documents), and ODP 46 (23,000 documents), and 20 Newsgroups dataset (19,996 documents), each with 17, 46 and 20 categories, respectively. We report the effectiveness in terms of precision, recall, and F1 measure.

A manual evaluation was conducted to identify relationships among categories. Three graduate students, familiar with the domain, participated in this evaluation. We only used the relationships that the majority of the human assessors agreed upon, with the average Pearson's correlation of 82% between each pair of the evaluators.

Our results (Figure 2) indicate that with relatively high recall (91%, 80% and 91%) we discover relationships among non-sibling categories for 20 Newsgroups, ODP17 and ODP 46, respectively, using *misclassification based* approach. However the precision of discovering these relationships is low, namely: 50%, 57% and 58% for 20 Newsgroups, ODP17 and ODP46 datasets, respectively. The F1 measure for finding non sibling relationships is 65%, 67% and 71% for 20 Newsgroups, ODP17 and ODP46, respectively. Using association rule mining to identify the relationships did not yield comparable results to those based on misclassification information.

Our results indicate that the proposed approach discovers most of the non-sibling relationships in a hierarchy. Using a higher weight threshold on the links (relationships), the lower weight links can be eliminated to improve the precision.

4. CASE STUDY

We present two case studies from our results. In Figures 3 and 4, we illustrate *networked hierarchies* for ODP17 and 20 Newsgroups datasets, respectively. The non-hierarchical links indicate relationships that are discovered among non-sibling categories (nodes). The solid-line links are the non-sibling relationships identified during manual evaluation (true positives), and dashed-line links indicate the relationships that were not identified during manual evaluation (false positives).

Our approach correctly discovered relationships such as <Baseball, Entertainment>, <Football, Games>, <Games, Entertainment>, <Games, Graphics> and <Hacking, Crime> in the ODP 17 dataset and <Atheism, Christian>, <Christian, Politics-misc>, <Religion-misc, Christian>, <Atheism, Politics-misc>, <Atheism, Religion-misc>, <Pc-hardware, Sci-electronics>, etc. in the 20 Newsgroups dataset. These relationships were not evident in the original hierarchies. Furthermore, unlike the *Relationship-net* that lost some of the hierarchical structure between parent-child and sibling, the *networked hierarchy* maintained all the hierarchical structure and relationships.

In summary, we proposed a new structure called *networked hierarchy* that represents category relationships that unlike a network such as relationship-net preserves the hierarchy of the categories and unlike a hierarchy represents also the non-sibling relationships among categories.

5. REFERENCES

[1] S. Mengle, N. Goharian, Detecting Relationships among Categories using Text Classification, JASIST , 61 (5), 2010

[2] S. Mengle, N. Goharian, Mining Temporal Relationships Among Categories, ACM 25th Symposium on Applied Computing, 2010

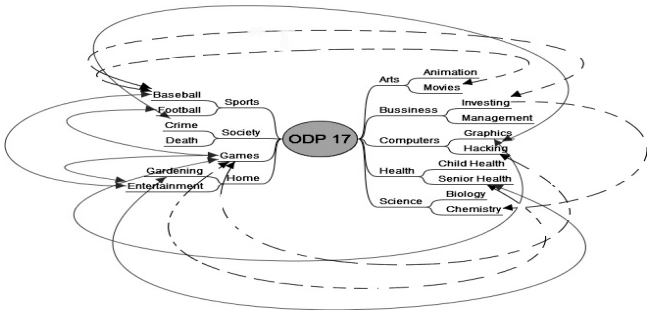

Figure 3. Case study for ODP 17 dataset

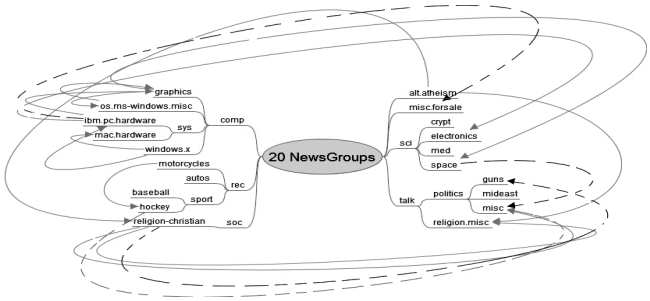

Figure 4. Case study for 20 Newsgroups dataset