Discovering Relationships among Categories using Misclassification Information

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

Nazli Goharian
Information Retrieval Lab
Computer Science Department
Illinois Institute of Technology
Chicago, Illinois, U.S.A
nazli@ir.iit.edu

Alana Platt
Information Retrieval Lab
Computer Science Department
Illinois Institute of Technology
Chicago, Illinois, U.S.A
platt@ir.iit.edu

ABSTRACT

Knowledge of relationships among categories is of the interest in different domains such as text classification, content analysis, and text mining. We propose and evaluate approaches to effectively identify relationships among document categories. Our proposed novel method capitalizes on the misclassification results of a text classifier to identify potential relationships among categories. We demonstrate that our system detects such relationships, even those relationships that assessors failed to identify in manual evaluation. Furthermore, we favorably compare the effectiveness of our methods with the state of art method and demonstrate a significant improvement in precision (34%) and recall (5%).

Categories and Subject Descriptors

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

General Terms

Algorithms, Performance, Experimentation

Keywords

Category relationships, Classification, Clustering

1. Introduction

Text classification algorithms automatically assign a predefined category to a given document. Most of these methods assume that all the categories are independent and mutually exclusive from each other. The reality is that in many domains, the categories are neither conditionally independent from each other, nor mutually exclusive [12]. The relationships that exist among categories depend on the contents of the documents that are in that category. The goal of our research is to find these relationships among categories based on the documents that belong to each category. Next, we provide several application and examples that motivate our research.

- Knowledge of relationships among different categories has shown to improve the effectiveness of SVM text classifiers [12], as well as the effectiveness of the standard Naïve Bayes model [8].
- Recent efforts such as [13] use the knowledge of closeness among different categories in the Open Directory Project (ODP) [5] and Yahoo Directories [11] to compute the closeness between instances of concepts, i.e., names of artists, locations, brand names, etc. This knowledge is useful in linguistics, the semantic web, and also in text mining.
- By discovering the relationship between two categories, one can provide cross-links from one category to another assuming that the customers looking at one category may be interested in buying the product in other category too. For example, by discovering a relationship between category *digital cameras* and *memory cards*, one can automatically provide a cross-link from web pages that sell cameras to web pages about memory cards.
- Similarly, another application may be in the newspaper and blog websites. For example, if a person is reading a news article about *movie reviews*, then he/she may also be interested in news articles about *Hollywood gossip* rather than news about *medical discoveries.*
- Knowledge about relationship among categories is also useful for analysts in fields of security, medical, or business. Finding similarity between categories may assist the analysts to understand a problem better, and hence, discover various issues related to that problem.

As shown in the above examples, the knowledge of relationships among categories is useful in different applications. Various manual efforts like Yahoo Directories and Open Directory Project aim to create a hierarchy of categories. Although these efforts are successful to an extent, there are several drawbacks. Firstly, the categories are constructed manually. Hence, as the number of categories increases, the amount of manual effort needed also increases. Secondly, human judgments are not only prone to errors, but also may be based on a limited knowledge on a particular topic. Thus, expert knowledge is needed to correctly construct hierarchy of categories related to different domains. Thirdly, there may be a relationship between two categories that may not have the same parent. For example, in ODP dataset, the contents of category *Animation* and *Computer Graphics* are closely related to each other. However, *Animation* is listed under *Arts* category while *Computer Graphics* is listed under *Computers* category. Such relationships among categories belonging to branches of different sub-trees cannot be represented in the taxonomy of categories. Finally, there may also be a relationship that does not seem obvious to a human editor although it exists. In

SAC'08, March 16-20, 2008, Fortaleza, Ceará, Brazil.

the later part of this paper, we provide a case study that presents an example of such a case.

We propose a novel method that uses the misclassification information obtained from a text classifier to find relationships among categories. We hypothesize that most of the misclassifications occur in the categories that are indeed related to their actual corresponding categories.

2. Prior Work

A large scale manual effort is being done to create a hierarchy of categories where similar categories are grouped and kept under a root. The Yahoo Directories has 292,216 categories and ODP has 118,488 categories [3,5,11]. Projects like Yahoo Directories and Open Directory Project (ODP) utilize a large amount of manual effort to define the structure of their category tree.

Limited efforts have been done to automatically find the relationships among categories [3,6,10]. In [10], a method named divide by two (DB2) is discussed, where categories are recursively subdivided into two groups until each group has only one category. The classes are recursively divided into two groups based on the class mean distance from the origin. This method ensures that a near perfect binary tree is generated which improves the time complexity of multiclass SVM classifier as compared to one-against-rest and one-against-one methods. In [10], only the efficiency of the approach is discussed and not the effectiveness in finding relationships among categories. Similar effort is discussed in [6] where the categories are recursively subdivided to create a tree. In every recursion, the two categories whose mean vectors are farthest from each other are selected as centroids and the remaining categories are assigned into these two clusters using spherical K-means clustering algorithm.

In [3], a bipartite graph is used to map the categories and documents and another bipartite graph is used to represent relationships between documents and terms. Single value decomposition is used to partition the two bipartite graphs and generate a taxonomy. [3] compares the tree generated by biparte graph method with the natural hierarchical structure of 20 News Group. As their goal was to demonstrate that automatically generated hierarchy of categories can be used by a hierarchical multiclass SVM model, the results for multiclass SVM were presented. Natural taxonomy of 20 News Group dataset yielded better results than the method proposed in [3] when used for multiclass hierarchical SVM. Thus, if the relationship between two classes is not predicted correctly, it may induce errors in multiclass SVM. Hence, we need a method that predicts relationships with a high precision so that wrong relationships are not predicted. We compare our results with the results reported in [3] in Section 6.

3. Methodology

Our aim is to discover relationships among categories while keeping the false positives (wrongly identified relationships) and false negatives (unidentified relationships) at its minimum. To solve this problem, we propose an approach that uses the misclassification information after the documents are classified/categorized into apriori known categories. We also are interested to examine the effectiveness of a rather simplistic approach that is based on calculating similarities between categories to achieve the same goal. In Section 3.1 and 3.2, we present *distance (similarity)* based approach and *misclassification based* approach, respectively.

3.1 Distance Based Approach

Similarities among categories are used to discover relationships among them. Each document is originally mapped to a category (or more). Thus, documents falling under the same category are grouped together. This process is called segmentation which is similar to clustering [7]. Using Euclidean distance, we find the closeness among the groups. The formula to calculate the distance between category *i* and category *j* is as follows:

$$d_{i,j} = \sqrt{\sum_{k=1}^{m} (tf_{ik} - tf_{jk})^2}$$

where tf_{ik} is term frequency of a term *k* in category *i* and *m* is the total number of unique terms in the union of *i* and *j*.

We find the *n* closest categories to a given category and define that as the predicted relationship(s). Empirically we determined that using only one closest category leads to higher precision results while considering top two categories leads to a better recall. When *n* is set to a value more than two, the performance of the system degrades drastically. The results for distance based approach are given in Section 6.1.

3.2 Using Misclassification Information

The premise is to use the misclassification information to find interesting relationships among categories. Our observation indicates that many misclassifications occur due to the existing relationships among categories. Table 1 shows an example of confusion matrix generated for 20 Newsgroup dataset using FACT statistical text classifier [4]. For this illustration, only six categories are shown for readability. Each column of the confusion matrix represents the instances in a predicted category, while each row represents the instances in an actual category. The heirarchy of 20 Newsgroups dataset (Figure 1 in Section 4) shows that categories like *Baseball and Hockey* are listed under the parent *Sports*; *PC and Mac* are listed under the parent *Sys* and categories *Autos* and *Motorcycle* are listed under the parent *Rec*. The shaded areas in Table 1 demonstrate that most of the misclassifications occur among the related categories rather than unrelated categories. This information is useful and may help in finding the closeness (similarity) between two categories. This observation motivates our approach in using misclassification information to discover relationships between two categories.

Finding the relationships among categories may be the pre-processing step for many applications. The FACT classifier trains and tests in linear time. Thus we used FACT as our text classifier.

Table 1: Confusion Matrix for subset of 20 News Group

		Predicted Category					
Actual Category		**B**	**H**	**P**	**M**	**AU**	**MO**
	B (Baseball)	405	7	1	1	3	4
	H (Hockey)	5	527	0	1	1	0
	P (PC)	2	1	171	17	7	4
	M (Mac)	1	3	13	243	8	7
	AU (Autos)	1	1	4	6	453	9
	MO (Motorcycles)	1	0	0	2	19	505

Our future plans include testing the misclassification algorithm on different text classifiers to observe any potential changes in our findings.

During the training phase of the text classifier, FACT assigns each term an ambiguity measure that indicates how unambiguous a term is with respect to a given category. Thus, the most unambiguous terms are retained during the training phase. During the testing phase, we use these unambiguous terms to determine the category of a document. If the term consistently (threshold of 60% was shown to be best in [4] using FACT) appears in a given category during the training phase and appears in different category during the testing phase, it indicates that the term may be a common keyword in both the categories and thus, may lead to misclassifications. The more common keywords among categories, there is higher the rate of misclassification. Thus, we are interested to evaluate our hypothesis that "the category that is predicted the most number of times and is different than the actual category may indeed be related to the actual category". The other methods discussed in the prior work section, as well as the distance based method, do not consider the knowledge about each term that is obtained during the training phase of a classifier. Hence, our method using misclassification information to discover relationships among categories significantly outperforms the existing methods. This is empherically shown in Section 6.

Our methodology for using misclassification to find relationships among categories follows four steps. A detailed explanation of each step is given below. We use the example of the confusion matrix for subset of 20 News Group dataset, shown in Table 1, to explain our methodology. Details about 20 News Group dataset are given in Section 4.

Step 1: Nullifying the effect of true positives

Our focus is on the misclassified documents and information pertaining to them. Thus, the correct predictions, i.e., true positives are nullified by setting them to zero (Table 2).

$$M(j,k) = 0 \; if \; j = k \qquad ..3.1$$

where M is the matrix of Table 2 with rows *j* and columns *k.*

Step 2: Pre-processing data by normalization

The number of training documents in different categories varies. Some categories have a large number of training documents and thus, a higher probability of having more keywords. Thus, they tend to be predicted more often than other categories that have less training documents. For example (Table 2), category *Autos* is predicted more (38 times) than the category *Motorcycles* (24 times). We have to normalize the values such that

Table 2: Confusion Matrix M for subset of 20 News Group with all the correct classifications (true positives) set to zero

		Predicted Category (Misclassified)					
		B	**H**	**P**	**M**	**AU**	**MO**
Actual Category	**B (Baseball)**	0	7	1	1	3	4
	H (Hockey)	5	0	0	1	1	0
	P (PC)	2	1	0	17	7	4
	M (Mac)	1	3	13	0	8	7
	AU (Autos)	1	1	4	6	0	9
	MO (Motorcycles)	1	0	0	2	19	0

misclassifications occurring when *Autos* is predicted are comparable with the number of misclassifications when category *Motorcycle* is predicted. All the values are normalized for a given predicted category using the sum of the number of times that the predicted category was misclassified. This allows the normalization of all the values in the range [0–1]. We use Formula 3.2 to normalize the values. The confusion matrix shown in Table 3 shows the normalized values for each category.

$$M_N(j,k) = \frac{M(j,k)}{\sum_{i=1}^{n} M(i,k)} \qquad ..3.2$$

where *j* is the row and *k* is the column of matrix; *n* is the number of categories; and $1<=j<=n$ and $1<=k<=n$.

Step 3: Measuring the category similarities

Based on the normalized matrix, we find the category ($C_{maxFN(j)}$) that has the highest number of false negative (the situation when a document is incorrectly classified as category C_k when the actual category is C_j and $j \neq k$).

$$C_{maxFN(j)} = \{C_k \mid \max(M_N(j,k)) \; for \; j \neq k\} \qquad ..3.3$$

As an example, consider the category *Baseball*. The highlighted row shows the normalized values of false negatives when the actual category is *Baseball*. Based on that, we find that the category *Hockey (0.58*) has the highest false negative value. Thus, $C_{maxFN(Baseball)}$ is set as *Hockey.*

Based on the normalized matrix, we also find the category ($C_{maxFP(j)}$) which has the highest number of false positive (the situation when the actual category of a document is C_k and the document is predicted as C_j and j≠k).

$$C_{maxFP(j)} = \{C_k \mid \max(M_N(k,j)) \; for \; j \neq k\} \qquad .. 3.4$$

As an example, the highlighted column in Table 3 shows the normalized false positives values when the category was predicted as *Baseball.* Based on that, we find that when *Baseball* is predicted, category *Hockey* has the highest value (0.50) for false positives. Thus, we set $C_{maxFP(Baseball)}$ as *Hockey*. We calculate $C_{maxFP(j)}$ and $C_{maxFN(j)}$ for all the categories and store them in a table called *Category Relation* as shown in Table 4.

Table 3: Normalized confusion matrix M_N for subset of 20 News Group

		Predicted Category (Misclassified)					
		B	**H**	**P**	**M**	**AU**	**MO**
Actual Category	**B (Baseball)**	0.00	**0.58**	0.06	0.04	0.08	0.17
	H (Hockey)	**0.50**	0.00	0.00	0.04	0.03	0.00
	P (PC)	0.20	0.08	0.00	0.63	0.18	0.17
	M (Mac)	0.10	0.25	0.72	0.00	0.21	0.29
	AU (Autos)	0.10	0.08	0.22	0.22	0.00	0.38
	MO (Motorcycles)	0.10	0.00	0.00	0.07	0.50	0.00

Step 4: Predicting relationships between categories

For a given category C_j, when $C_{maxFP(j)} = C_{maxFN(j)}$, we predict that a *strong relationship (SR)* exists between C_j and $C_{maxFP(j)}$ (or $C_{maxFN(j)}$). A *strong relationship (SR)* between category C_j and C_k is defined as:

$$SR(C_j, C_k) \; if \; ((C_k = C_{maxFP(j)}) \; and \; (C_k = C_{maxFN(j)})) \qquad ..3.4$$

For the subset of 20 News Group dataset, we predict that *strong relationship* exists between *Baseball* and *Hockey;* between *PC* and *Mac,* and between *Autos* and *Motorcycles*. If $C_{maxFN(j)}$ is not equal to $C_{maxFP(j)}$, we predict that a *weak relationship (WR)* exists between C_j and $C_{maxFN(j)}$ and C_j and $C_{maxFP(j)}$. A *weak relationship (WR)* between categories C_j and C_k is defined as:

$$WR(C_j, C_k)\ if\ ((C_k = C_{maxFP(j)}\ or\ C_k = C_{maxFN(j)})\ and\ (C_{maxFN(j)} \neq C_{maxFP(j)})) \quad ..3.5$$

We evaluate two scenarios where we (a) predict only the strong relationships and (b) both strong and weak relationships. Predicting *only strong relationships* is useful in situations where precision of prediction is more important than recall. Predicting both *strong and weak relationship* is useful when both precision and recall are equally important. Results for approach using misclassification information are given in Section 6.1.

Table 4: Category Relation for subset of 20 News Group

Category(C_j)	$C_{maxFN(j)}$	$C_{maxFP(j)}$
Baseball	Hockey	Hockey
Hockey	Baseball	Baseball
PC	Mac	Mac
Mac	PC	PC
Autos	Motorcycles	Motorcycles
Motorcycles	Autos	Autos

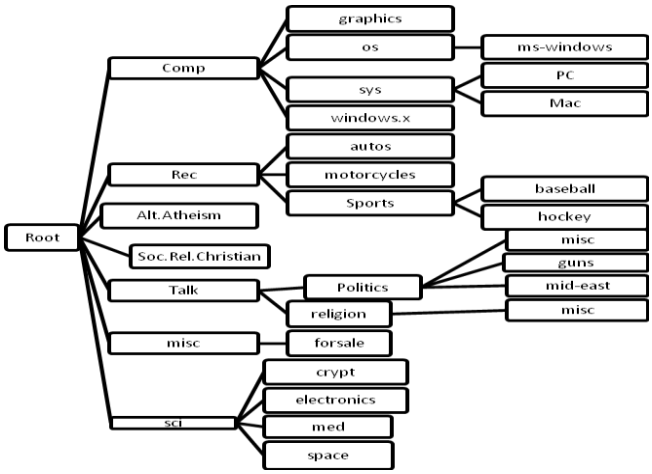
4. Datasets

We use 20 Newsgroup and Open Directory Project datasets for experimentations. These datasets are commonly used benchmark datasets in the field of text classification.

20 News Groups dataset

The 20 News Groups dataset [1] is already divided into apriori known categories. It consists of 20,000 documents categorized into 20 different categories. Each category has 1000 documents assigned to it. The hierarchy of 20 newsgroups dataset is as given in Figure 1.

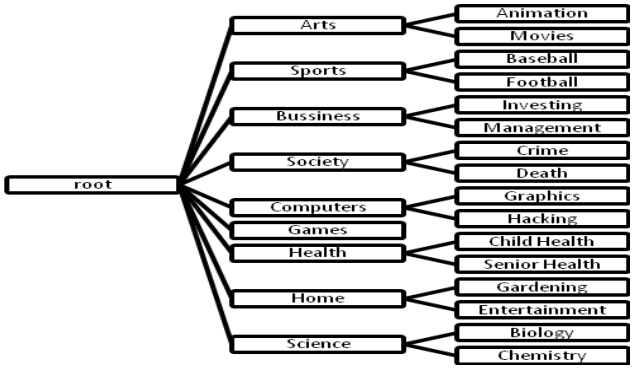
Figure 1: Hierarchy of 20 Newsgroups



Open Directory Project dataset

The Open Directory Project (ODP) [5] is a comprehensive human edited directory of the Web, compiled by a vast global community of volunteer editors. It consists of a pre-defined hierarchy of categories with links associated with each category. We select a subset of ODP with 17 categories and 500 documents per category. The categories and its hierarchies are illustrated in Figure 2. The categories belong to different domains. The children of the same parent categories in the tree are closely related to each other.

Figure 2: Hierarchy of subset of ODP dataset



5. Evaluation

5.1 Evaluation metrics

Precision is defined as the measure of how accurately relationships among categories are predicted without predicting the relationships that do not exist. Precision is defined using the Formula 5.1.

$$\text{Precision(P)} = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad .. 5.1$$

Recall is defined as the ratio of correctly identified relationships to total existing relationships. The undetected relationships are false negatives. Recall is defined using Formula 5.2.

$$\text{Recall(R)} = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad .. 5.2$$

We also use F1 measure which combines both Precision (P) and Recall (R) with an equal importance. F1 measure is defined using the Formula 5.3.

$$\text{F1measure} = \frac{2.P.R}{(P+R)} \quad .. 5.3$$

5.2 Manual Evaluation

To validate our approach, we use human evaluations to identify relationships among the categories in each of the datasets. Five graduate students participated in this evaluation. Each of the human assessors was explained apriori about the content of each category and a summary of documents in each category was given. Subsequently, the evaluators each individually tagged relationships among categories using three different levels:

Level 1: No relationship exists
Level 2: Relationship may exist
Level 3: Strong relationship exists

We only use the relationships that majority of the human assessors marked as *Level 3.* The average Pearson's correlation between

each pair of the evaluators was 86.18%. Results of the human evaluations for both datasets are presented in Table 5 and Table 6.

Table 5: Relationships identified by human evaluators in 20 NewsGroup dataset (The two categories in each row are tagged as relationships by the human evaluators)

Categories having the same parents		
comp.sys.ibm.pc.hardware	⇔	comp.sys.mac.hardware
rec.auto	⇔	rec.motercycles
rec.sports.baseball	⇔	rec.sports.hockey
talk.politics.guns	⇔	talk.politics.mideast
talk.politics.guns	⇔	talk.politics.misc
talk.politics.mideast	⇔	talk.politics.misc
comp.graphics	⇔	comp.windows.x
alt.atheism	⇔	soc.religion.christian
Categories having different parents		
alt.atheism	⇔	talk.religion.misc
soc.religion.christian	⇔	talk.religion.misc
comp.graphics	⇔	comp.os.ms-windows.misc
comp.graphics	⇔	comp.sys.ibm.pc.hardware
comp.graphics	⇔	comp.sys.mac.hardware
comp.os.ms-windows.misc	⇔	comp.sys.ibm.pc.hardware
comp.os.ms-windows.misc	⇔	comp.windows.x
comp.sys.mac.hardware	⇔	comp.windows.x
comp.sys.mac.hardware	⇔	sci.electronics

Table 6: Relationships identified by human evaluators in ODP

Categories having the same parents		
Arts/Animation	⇔	Arts/Movies
Business/Investing	⇔	Business/ Management
Computer/Graphics	⇔	Computer/Hacking
Health/Child Health	⇔	Health/Senior Health
Home/Gardening	⇔	Home/Entertainment
Science/Biology	⇔	Science/Chemistry
Society/Crime	⇔	Society/Death
Sports/Baseball	⇔	Sports/Tennis
Categories having different parents		
Arts/Animation	⇔	Computer/Graphics
Arts/Movies	⇔	Home/Entertainment
Games	⇔	Arts/Animation
Computer/Graphics	⇔	Games
Arts/Movies	⇔	Computer/Graphics
Computer/Hacking	⇔	Society/Crime

6. Results

The results section is organized into three subsections. In section 6.1, we present the effectiveness of our *distance based* and *misclassification* approaches using the two benchmark datasets, i.e., Open Directory Project and 20 Newsgroups. In Section 6.2, we compare our approach using misclassification information with the state of the art method using the same dataset. Section 6.3 evaluates our approach in identifying relationships among categories with the same parents versus with different parents.

6.1 Distance based vs. Misclassification Information

We evaluate our two approaches, i.e., *distance based* and *misclassification information*.

In the *distance based approach*, relationships between a given category and *n* closest categories are predicted. As observed in Table 7, when only the relationship between a category and its closest category (*Dist N1*) is predicted we obtain a better precision

Table 7: Results of distance based approach and approach using misclassification information based on 20 NG and ODP dataset

Methods	**20 News Group**			**Open Directory Project**		
	Pre.	**Recall**	**F1**	**Pre.**	**Recall**	**F1**
Dist N1	0.64	0.55	0.59	0.35	0.42	0.38
Dist N2	0.44	0.65	0.52	0.30	0.71	0.42
Misclass SR	**0.88**	0.40	0.55	**0.75**	0.18	0.29
Misclass SWR	0.59	**0.68**	**0.63**	0.61	**0.84**	**0.71**

(20NG: 0.64, ODP: 0.35) than the situation where we identify the two closest categories (*Dist N2*) with respect to a given category (20NG: 0.44, ODP: 0.30). This is because the probability of the existence of a relationship between the closest category to a category is more than the second closest category. As we try to identify more potential related categories to a given category, precision decreases while recall increases. When we predict relationships between a given category and the two closest categories, the recall increases both in the case of ODP dataset (*Dist N1=0.42, Dist N2=0.71*) and 20 News Group dataset (*Dist N1=0.55, Dist N2=0.65*).

In the *misclassification information* approach, we evaluate two methods. In the first method (Misclass SR), only *strong relationships* that exist among categories are predicted. As observed in Table 7, using this method always gives the highest precision value *(20NG: 0.88, ODP: 0.75)* on both datasets. Hence, using the *misclassification information* and *strong relationships* ensures that only the closest relationships are predicted.

In method *Misclass SWR*, where we predict both *strong and weak relationships*, precision decreases *(20NG: 0.59, ODP: 0.61)* while the recall increases *(20NG: 0.68, ODP: 0.84).* As observed in Table 7, using misclassification information and predicting both strong and weak relationships (*Misclass SWR*) finds the maximum number of relationships among categories.

6.2 Comparison with Related work

To our knowledge, limited work has been done in the field of finding relationships among categories. Thus, we compare our proposed approach based on the *misclassification information* with the state of the art effort called *Consistent Biparte Spectral Graph Co-partition (CBSGC)* [3]. As no evaluation metrics were used to show that the relationships which were found using CBSGC algorithm are valid, we use precision and recall as our evaluation measures to compare the results of CBSCG with our proposed approach of *misclassification information.*

As shown in Table 8, both variations of the proposed approach of *misclassification information* (*Misclass SR and Misclass SWR*) outperform *Consistent biparte spectral graph co-partition*. Misclass SR method that predicts only the strong relationships among categories performs the best in terms of precision (0.88). However, Misclass SWR method that predicts both strong and weak relationships among categories performs best in terms of recall (0.68) and F1 measure (0.63).

Table 8: Comparison of our proposed approach of misclassification information and CBSCG on 20 News Group dataset

Method	**Precision**	**Recall**	**F1**
Misclass SR	**0.88**	0.40	0.55
Misclass SWR	0.59	**0.68**	**0.63**
CBSGC	0.54	0.35	0.42

6.3 Discovering the relationships among categories with same parents vs. relationships among categories with different parents

To demonstrate how our proposed method performs in discovering relationships among categories that have different parents versus same parents in the hierarchy of categories, we present the corresponding results in Table 9. It can be observed that *Misclass SR* only finds relationships within same parents. This is because strong relationships mostly exist between children of the same parent class. However, using *Misclass SWR*, we can also find relationships among categories with different parents. Thus, it can be noted that, if we already have knowledge about the hierarchy of categories, we can still apply *Misclass SWR* to find the hidden relationships that are not visible in hierarchy of categories but present in the dataset.

Table 9: Results of our approach of misclassification information with respect to categories with same parents vs. different parents

		Same Parent			Different Parents		
Dataset	**Method**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
ODP	Misclass SR	0.75	0.18	0.29	0.00	0.00	0.00
	Misclass SWR	0.39	0.53	0.45	0.22	0.31	0.25
20NG	Misclass SR	0.88	0.40	0.55	0.00	0.00	0.00
	Misclass SWR	0.36	0.45	0.40	0.23	0.23	0.23

7. Case Study

We discuss one example from 20 newsgroups datasets, where the human assessors did not identify a relationship between two given categories and our algorithm detected such relationship. The category *comp.sys.ibm.pchardware* contains news reports about different issues related to IBM PC hardware. The category *misc.forsale* deals with news articles and advertisements regarding selling different items. The human evaluators did not tag a relationship between *comp.sys.ibm.pchardware* and *misc.forsale*. When we investigated the documents that were misclassified between categories *comp.sys.ibm.pchardware* and *misc.forsale*, we noticed that for 20 News group dataset, most of the documents that fell under *misc.forsale* category and were misclassified as *comp.sys.ibm.pchardware* dealt with issues regarding PC sale and computers configuration. This does not imply that a close relationship exists between *comp.sys.ibm.pchardware* and *misc.forsale* in general, but for the specific 20NG dataset, a relationship indeed exists between *comp.sys.ibm.pchardware* and *misc.forsale*.

8. Conclusion

Effectively discovering relationships among categories is useful in the field of text mining and text classification. We proposed an approach that uses the misclassification information to find relationships among the categories. We also proposed an approach that uses Euclidean distance between groups of documents of the same category and compared it with our misclassification method. We evaluated our proposed methods on 20 Newsgroup and Open Directory Project datasets, which are commonly used standard benchmark datasets in the field of text classification. Our experimental results showed that approaches using the misclassification information are more effective than distance based approaches.

To validate the effectiveness of our proposed approach using the misclassification information, we performed a comparison with the state of the art effort that finds relationships among the categories. It was found that the proposed approach using the misclassification information outperforms the consistent biparte spectral graph co-partition method. We proposed and evaluated that for a high precision system, using only strong relationships is the best. When we need a balanced system with a good recall and precision, using both strong and weak relationships was shown to be the best. Our analysis shows that *Misclass SR* only identifies relationships between categories with the same parents, while *Misclass SWR* also identifies relationships between categories that have different parents. We also showed that in some cases, our automatic algorithm was able to find relationships that human assessor were not able to identify.

9. References

[1] 20 News Groups dataset. http://people.csail.mit.edu/jrennie/20Newsgroups

[2] Dhillon I., Mallela S. and Modha D., Information-Theoretic Co-Clustering. The 9th ACM SIGKDD International Conference on Knowledge Discovery and Data mining, 2003

[3] 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 Transactions on Knowledge and Data Engineering Volume 17 Issue 9, Special Issue on Data Preparation, 2005

[4] Mengle S, Goharian N and Platt A., FACT: Fast Algorithm for Categorizing Text. The 5th IEEE International Conference on Intelligence and Security Informatics, 2007

[5] Open Directory Project (http://dmoz.org)

[6] Punera K., Rajan S. and Ghosh J., Automatically Learning document taxonomy for Hierarchical classification. The 14th International Conference on World Wide Web, 2005

[7] Tan P.N., Stienbach M, and Kumar V, Introduction to Data Mining, Addison Wesley, 2006

[8] Toutanova K., Chen F., Popat K. and Hofmann T., Text classification in a hierarchical mixture model for small training sets. The 10th ACM International Conference on Information and Knowledge Management, 2001

[9] Vapnik Vladimir, Statistical learning theory. Wiley, 1998

[10] Vural V. and Dy J., A Hierarchical Method for Multi-Class Support Vector Machines. The 21st International Conference on Machine Learning, 2004.

[11] Yahoo Directories (http://dir.yahoo.com)

[12] Zhu S., Ji X, Xu W. and Gong Y., Multilabelled Classification Using Maximum Entropy Method. The 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2005

[13] Ziegler C-N., Simon K. and Lausen G., Automatic computation of semantic proximity using taxonomy knowledge. The 15th ACM international Conference on Information and Knowledge Management, 2006