Cost Sensitive Evaluation Measures for F-term Patent Classification

Yaoyong Li, Kalina Bontcheva and Hamish Cunningham
Department of Computer Science, The University of Sheffield
211 Portobello Street, Sheffield, S1 4DP, UK
{yaoyong, kalina, hamish}@dcs.shef.ac.uk

Abstract

Some classification problems, such as the NTCIR-06 F-term patent classification task, have general or specific relations between the target class labels. Consequently, in such cases, it is desirable that the relations among the labels can be taken into account in the evaluation measures. For example, if a system assigns an incorrect label to one instance and the assigned label has close relation with the true label, then the system may deserve some credit, rather than being given no credit at all as is the case with conventional evaluation measures.

In this paper we propose some new evaluation measures based on relations among classification labels, which can be seen as the label relation sensitive version of some important evaluation measures such as averaged precision and F-measure. We also present the results of applying the new evaluation measures to all submitted runs for the NTCIR-6 F-term patent classification task. The new measure did change the ranking positions of some runs (including the top runs in some cases) made by the conventional measure, while it kept the ranking position of many other runs.

Keywords: *Evaluation Measure, Patent Classification, NTCIR, A-Precision.*

1 Introduction

In some multi-class classification problems, the class labels are related with each other in a certain fashion. One such example is patent classification based on the F-term, one subtask in NTCIR-5 and 6 evaluation schemes [9], in which the class label F-terms within one theme are related with each other in a tree structure (see Section 2.4 for further details). Another example is ontology-based information extraction, which is a multi-class classification problem where the class labels are concepts in an ontology (see, e.g. [12]). Hereafter we refer to this kind of problems as *hierarchical classification* problems.

In hierarchical classification problems, if an exact classification cannot be determined, then a label taken from the appropriate part of the class hierarchy is better than assigning a label randomly or not producing any label at all. For example, if a system cannot classify a new article about a football match correctly into category *Football*, it would make more sense to classify the article into category *Sport* rather than *Politics* in most cases.

Formally we can define a number $c(X, Y)$ to measure the misclassification cost of classifying an instance of class X into class Y. A system's performance measure should be related to the cost — the less cost the system obtains on evaluation data, the higher performance the system has. Clearly, if the class labels have the specific/general relations with each other, it would make sense to make the cost $c(X, Y)$ rely on the relation of the class X and Y. The closer relation the two classes have, the lower the misclassification cost between them should be.

However, unfortunately, the conventional measures, such as those used in the F-term patent classification subtasks in NTCIR-5 and NTCIR-6, simply defined the cost as a binary value, which is equal to 1 if the two classes are the same and 0 otherwise. Obviously binary cost does not take into account the relations among the classes[1].

It is worth noting that some previous work on hierarchical classification has used evaluation measures which were sensitive to the relations among the target classes. For example, [13] defined the cost of two classes as the height of the first common ancestor of the two classes in the taxonomy in their taxonomical document classification experiments, and [6] used the distance of two concepts in taxonomy as the cost. Both authors used the error rate based on the cost they defined to measure the results of their experiments. [12] defined cost for ontology-based information extraction, based on the BDM measure (which will be explained in more detail in Section 2) and used F-measure based on this cost to measure system per-

[1] It should be noted that the organisers of the patent classification subtask at NTCIR-6 had recognised the problem of using binary cost for the F-term classification and encouraged the participants of the subtask to consider new evaluation measures which take into account the relations among the F-terms, which is the main motivation of the work presented here.

formance.

However, none of this work has adopted the structure-sensitive cost in an important measure in information retrieval and document classification, such as averaged precision, which is one of the main contributions of this work.

In fact, averaged precision has been adopted as the primary measure in the NTCIR patent classification subtask and as one of primary measures in the information retrieval tracks of the TREC competition (see http://trec.nist.gov/pubs.html). However, both use averaged precision based on binary cost, which does not take into account the relations among the categories.

In this paper we present the averaged precision based on general cost. In particular, we adapt the averaged precision for using the BDM cost measure and use it to evaluate all submitted systems for the formal run of the patent classification subtask at NTCIR-6, which is another main contribution of this paper.

The rest of the paper is organised as follows. Section 2 discusses the definition of misclassification cost, in particular the one based on the BDM measure. It then describes averaged precision as well as the F-measure based on the general cost. Section 3 presents the results of using the BDM-based averaged precision on all runs submitted to the NTCIR-6 patent classification subtask, and compares them with the official results which are based on binary cost. Section 4 concludes with a summary and a discussion.

2 Cost-Based Evaluation Measures

2.1 Misclassification cost

Misclassification cost has been studied mainly in two communities: hierarchical classification learning and ontology engineering.

Work on hierarchical classification uses some non-binary cost in the learning algorithms as well as in defining cost-sensitive evaluation measures. Actually the advantage of a hierarchical classification algorithm over flat classification is often demonstrated better by the cost based measure than by the conventional binary measure. However, since the cost definition is not their main concern, the cost functions used are often very simple. For example, in both [6] and [3] where new margin-based learning algorithms for hierarchical classification are investigated, the cost $c(X, Y)$ is defined as the distance between the two nodes X and Y in the class label tree, namely the number of edges in the shortest path connecting nodes X and Y. [13] defines the cost of two classes as the height of the first common ancestor of the two classes in the taxonomy in their taxonomic document classification experiments.

In contrast, cost functions have been studied extensively in the ontology engineering community (see e.g. [5]). Here ontology refers to a form of knowledge representation in which the concepts and their relations in a given domain can be represented by a graph, e.g. the nodes in a graph correspond to the concepts of the domain and the links between the nodes represent the relations between them. In this field it is paramount to be able to measure the cost of misclassifying instances under the wrong ontology class or to be able to compare two ontology or errors in automatically learnt taxonomies (see e.g. [12, 2, 7]).

There are a number of requirements that need to be satisfied in the definition of a quantitive measure between two concepts. In the following we first discuss some of those requirements. Then we introduce two promising cost functions, arising from ontology engineering, which are based on learning accuracy and the BDM, respectively.

[1] studied word sense disambiguation using WordNet, which can be regarded as an English lexicalised ontology. The paper suggested some criteria for measuring closeness of two concepts organised in a graph, which are useful for the definition of cost function between them:

1. The measure should be dependent on length of the shortest path connecting the two concepts involved.

2. The concepts in a deeper part of the hierarchy should be closer.

3. Concepts in a dense part of the hierarchy should be relatively closer than those in sparse region.

4. The measure should be independent of the number of concepts in the graph.

It is worth noting that the definition of cost function based on the shortest distance of two concepts in a graph meets the first criterion but not the others, and the cost definition used in [13] satisfies the second criterion only.

2.2 Learning accuracy

[4] use a measure called Learning Accuracy to assess how well an instance was added to ontology. This measure was used originally in [8] to measure whether a concept had been added at the right level of the ontology, but it can be equally applied to measure how well an instance has been added in the right place. Learning Accuracy (LA) essentially measures "the degree to which the system correctly predicts the concept class which subsumes the target concept to be learned".

LA uses the following measurements:

- SP = the shortest length from root to the key concept

- FP = shortest length from root to the predicted concept.
- CP = shortest length from root to the MSCA (Most Specific Common Abstraction, i.e. the lowest concept common to SP and FP paths)
- DP = shortest length from MSCA to predicted concept

The LA is defined via the following equation,

$$LA = \begin{cases} CP/SP = 1 & \text{if } FP = 0 \\ CP/(FP + DP) & \text{otherwise} \end{cases}$$

Essentially, this measure provides a score on a scale between 0 and 1 for any concepts identified in an incorrect position in the ontology. If a concept is missing or spurious, the score is 0, and if it is correct, the score is 1 (as with Precision and Recall). So this method provides an indication of how serious the error is, and weights it accordingly. We can deduce a cost from LA easily, e.g. as $1 - LA$.

We can see that LA is dependent on the shortest path connecting the two concepts and also the deepness of the two concepts in the ontology. Hence it is compatible with the first two criteria listed above. However, it does not take into account the local concept density. It is not normalised with respect to the size of ontology either. In another word, it does not meet the other two criteria.

Also from the definition we can see that LA is asymmetric for the two concepts involved, namely the key concept and the predicted concept, as DP is involved in the main definition of the LA but SP is not. It means that LA does not take into account the distance of the key concept from the MSCA. LA is equal to 1 as long as the predicted concept is an ancestor of the correct concept. In the extreme case, a useless classifier that always predicted the root concept would obtain a perfect LA score. Interestingly, an opposite intuition was adopted in [3] which stated that "if a mistake is made at node i, then further mistakes made in the subtree rooted at i are unimportant", meaning that it did not consider the distance between the predicted concept and the MSCA. In contrast, the BDM formula treats the two involved concepts symmetrically, as can be seen below.

2.3 BDM measure

Recently a new cost measure, called BDM, has been proposed for ontology-based information extraction [11, 12], which can be seen as an improved version of LA.

In detail, given a key concept K and a predicted concept R in an ontology, the BDM measure for K and R, $BDM(K, R)$, is defined as

$$\frac{CP/n_0}{CP/n_0 + DPK/(n_2 * BR) + DPR/(n_3 * BR)} \quad (1)$$

The parameters and variable used in above equation are explained in the following,

- CP is the length of the shortest path from the root concept to MSCA, as in the definition of LA.
- DPK and DPR are the lengths of the shortest paths from MSCA to the key and response nodes, respectively.
- n_2 and n_3 are the averaged lengths of chains (from the root node to a leaf node) containing the key and response nodes, respectively.
- n_0 is the averaged length of all chains (also from the root node to a leaf node) in the ontology graph, which is used in the formula for normalising the two specific chain lengths n_2 and n_3 such that the measure is not sensitive to the size of the ontology (refer to the fourth criterion).
- BR represents the concept density of the local area containing the key and response concepts, which is computed as the averaged number of branches of the nodes between the MSCA node and the key node or between the MSCA node and the response node, and is normalised by the averaged number of branches over all nodes in the graph.

Note that n_0, n_2 and n_3 are used together for representing the vertical density of the local area containing the key and predicted nodes. BR is used for measuring the traversal density of the local area. The larger BR results in the higher BDM score, which makes the BDM satisfy the fourth requirement listed above.

Finally, we may define a misclassification cost, based on the BDM measure as $C_{BDM}(R, K) = 1 - BDM(R, K)$, because the BDM measure is between 0 and 1 and is in proportion to the closeness of two nodes in the graph.

In comparison to the LA, BDM is normalised with respect to the size of ontology and also takes into account the concept density of the area containing the two involved concepts. Therefore the BDM satisfies all four criteria listed above. Moreover, BDM treats the key and predicted concepts equally, which is required in applications such as F-term patent classification.

2.4 Cost based evaluation measures

The three main measures used in the F-term patent classification subtask of NTCIR-6 are *A-Precision*, *R-Precision* and *F-measure*, of which the A-Precision is

the primary one. The official evaluation script of the subtask adopted a binary cost which is 1 for one exact match and 0 otherwise. Here we propose the extension of the three measures to a general cost function.

First note that the central part of the computations of all three measures is the computation of precision for any set of test examples. Both A-Precision and R-Precision are derived from a ranked sequence of text examples. A-Precision is the mean of the precisions at all recall levels of the ranked example sequence. R-Precision is the mean of the precisions at some predefined recall levels. F-measure is the harmonic mean of precision and recall for a particular subset of test examples that are defined by the system for evaluation.

Given a particular class X of a multi-class classification problem and a subset S of test example with the predicted class labels, the conventional precision for one class is the ratio of the number of the exactly matched examples n_{match} to the total number of the examples in the subset S, namely

$$P = \frac{n_{match}}{|S|} \tag{2}$$

where the exactly matched example means that the example has the true class label X and is correctly classified into the class X.

If the classes in the classification problem are related with each other and a cost function $c(X, Y)$ is defined on the relations of the classes, then in the computation of the precision, we can consider not only the exactly matched examples, but also those examples which should have been classified as X but were classified as another class, related to X. We refer to these latter examples as *partially matched examples*.

The contribution of a partially matched example can be calculated on the basis of the cost between the correct class and the class assigned by the system. Formally, assume that the test examples which are in the subset S and belong to class X constitute the set S_0, and each example e in set S_0 is classified into the class Y_e, then the precision based on the cost function is defined as

$$P_{cost} = \frac{\sum_{e \in S_0} (1 - c(X, Y_e))}{|S|} \tag{3}$$

Hence, the lower the cost $c(X, Y_e)$ is, the more contribution the example e makes to the precision. In other words, the more similar classes X and Y_e are, the smaller the misclassification mistake is.

Once we have defined cost-based precision for any given subset of test examples, the calculation of A-Precision, R-Precision and F-measure is straightforward, as defined above.

However, the F-term patent classification subtask at NTCIR-6 is a harder multi-class document classification problem, because each document may belong to more than one class. Thus, in such cases, some further consideration is needed in the definition of cost-based precision, which will be discussed in the following section.

3 Cost Based Measures for F-term Patent Classification

3.1 F-term patent classification

In this subsection we give a brief introduction to the F-term patent classification problem; for further details see [9]. Patent classification is a necessary step of patent processing. The most widely used patent classification taxonomy is IPC. The Japanese Patent Office provides a two-level patent classification scheme. The first level denoted as FI is an extension of IPC, which refers to a set of themes in the patent. For example, the theme *2C088* is about "Pinball game machines (i.e., pachinko and the like)". Each theme has a collection of viewpoints for specifying possible aspects of the patent within the theme. Each viewpoint has a list of possible elements. Those viewpoints and the corresponding elements for each theme are encoded by the F-terms of the theme, which are the second level of that patent classification scheme. The theme *2C088* has the viewpoint *AA* for "Machine detail", the viewpoint *BA* for "Processing of pachinko ball", and the viewpoint *BB* for "Card systems". The viewpoint *AA* has the elements such as *AA01* for "Standard pachinko games (i.e., vertical pinball machines)" and *AA65* for "Special pachinko games". Hence, the F-terms under each theme have specific/general relations among them.

A subset of the F-terms for theme *2C088* and the relations between those F-terms is illustrated in Figure 1. Note that a root node is added as the parent node of all viewpoint F-terms. In our experiments the BDM measure between two F-terms under the same theme is computed on the structure shown in Figure 1. The BDM measure for some pairs of the F-terms shown in Figure 1 are presented in Table 1.

Each system participating in the NTCIR-6 F-term patent classification subtask was required, given a patent and a theme that the patent belongs to, to classify the patent into F-terms of the given theme. In other words, the system assigned some suitable F-terms to the patent for the given theme, which are used to compute the F-measure of the system.

The task also required each participating system to rank of up to 200 F-terms for each patent and theme, which are used for the computation of A-Precision and R-Precision.

A peculiarity of this patent classification task is that a patent may have more than one themes and under each theme may have many F-terms. This is different from normal document classification where each

Table 1. The BDM measure for some pairs of the F-terms shown in Fig. 1.

BDM	AA	BA	AA01	AA65	AA02	AA03
AA	1.00	0.00	0.64	0.61	0.63	0.47
BA	0.00	1.00	0.00	0.00	0.00	0.00
AA01	0.64	0.00	1.00	0.45	0.90	0.75
AA65	0.61	0.00	0.45	1.00	0.50	0.38
AA02	0.63	0.00	0.90	0.50	1.00	0.76
AA03	0.47	0.00	0.75	0.38	0.76	1.00

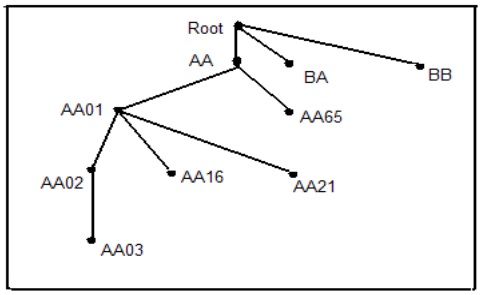

Figure 1. A subset of all F-terms for theme *2C088* and the relations between them

document is classified into one or more categories and the evaluation measure is computed for each category. Consequently, in F-term patent classification the evaluation measures for each document are computed from a ranked list or a subset of F-terms. In other words, the evaluation metrics for normal document classification use documents as the test examples while in F-term patent classification the test examples are F-terms. Therefore the cost-based measure for F-term classification is different from that for normal document classification and will be defined next.

3.2 Cost based measures for F-term classification

As already discussed, the evaluation of F-term classification is on a ranked list or subset of F-terms for each patent. In this case there are several ways in which one can calculate the partial matches in the precision formula, given a ranked list or a subset of F-terms (see 3). Here we consider two of these methods, one extremely generous and one extremely conservative.

First, let us assume that we want to compute a cost-based precision for a patent and the subset S_1 of F-terms which the system has assigned to the patent. The possibly most generous way is, for each F-term F in the considered subset, select the maximal BDM measure BDM_F among the BDMs between the F-term F and each of true F-terms for the patent. Then the precision of the subset S_1 for the patent is computed by

$$P_{BDM_{high}} = \frac{\sum_{F \in S_1} BDM_F}{|S_1|} \quad (4)$$

Note that in the above the BDM is used directly rather than the cost function.

Another method is the most conservative one. Assume that the F-term set T contains those F-terms in the set S_1 which are the true F-terms of the patent. For each F of those true F-terms for the patent which are not in the subset S_1, selects the maximal BDM measure BDM_F among the BDMs between the F-term F and those F-terms in the subset $S_1 - T$, namely the F-terms which are in the set S_1 but are not the true F-terms of the patent. Then the precision is computed by

$$P_{BDM_{low}} = \frac{\sum_{F \in (S_1 - T)} BDM_F + |T|}{|S_1|} \quad (5)$$

We used the BDM-based evaluation measure to evaluate all the submitted runs of the F-term patent classification subtask at NTCIR-6, which will be presented in next subsection. Since the precision computation (4) results in higher value than that by (5), for the sake of convenience, we refer to the BDM evaluation measure with (4) as BDM_{High} and the one based on (5) as BDM_{Low}.

3.3 Results Comparison

We have evaluated all 43 submitted formal runs (from different systems) for the F-term patent classification subtask at NTCIR-6, by using the evaluation measures based on *BDM_High* and *BDM_Low*, respectively. First, we computed the BDM scores for the F-terms under each of the themes considered. Then we re-used the evaluation script released by the organisers of the patent classification subtask at NTCIR-6, after making the necessary modifications on the part of the precision computations in the script by using the BDM measure and the precision computation formula (4) or (5) presented above. For the sake of simplicity, we refer to the officially released results as *Binary*, because they adopted a binary cost function.

The A-Precision results are presented in Figure 2 and Figure 3, which shows the absolute values of the BDM and binary scores and the differences in the values of these scores, respectively. The R-Precision results are in Figure 4 and Figure 5, whereas the F-measure ones are shown in Figure 6 and Figure 7.

In order to investigate the effects of the BDM scores on the ranking of the submitted runs, we computed the *Kendall tau rank correlation coefficient* for each of the BDM scores against the corresponding binary score (see e.g. [14]). The Kendall tau coefficient is used to measure the degree of correspondence between two rankings of the same objects. It has value of 1 if the two rankings are the same and -1 if one ranking is the reverse of another ranking. For other cases it has a value between -1 and 1. Particularly, if the two rankings are completely independent, it has a value of 0 For each of the three measures used in the NTCIR-6 patent classification subtask, e.g. A-Precision, we computed a Kendall's tau value for the two rankings of the 43 submitted runs respectively according to the binary score and one of the two BDM scores, by using an on-line Kendall's tau computation software [15].

Table 2 presents the Kendall's tau for the six pairs of rankings. All correlation coefficients are less than 1, showing that using the BDM scores did change the ranking of the submitted runs made by the binary measures. However, those coefficients were quite close to 1, in particular the values for A-Precision and R-Precision. Hence there are many agreements between the ranking produced by the BDM score and that produced by the binary score.

Table 2. The Kendall's tau for six pairs of rankings. Each pair consists of one ranking of all the submitted runs by one BDM score and another ranking by the corresponding binary score.

	A-Precision	R-Precision	F-measure
High	0.887	0.861	0.614
Low	0.912	0.848	0.752

We also manually checked the ranking changes made the BDM scores. For A-Precision, the 8 runs with the highest binary A-Precision scores did not change the ranking by any of the two BDM scores. For R-Precision, by using the high BDM score the run GATE-3 became the third highest, while it was the highest according to the binary score. However, the R-Precision scores of GATE-03 were very close to those of the two runs NCS01 and NCS02 which were swapped with the GATE-03, as shown in Table 3. For the F-measure, GATE-04 had the same high BDM score as NCS02, but the binary score of the former was slightly lower than that of the latter (the F-measures of both were only lower than that of GATE-03). Therefore, in some cases the BDM scores changed the ranking of the top runs, but only because those runs had very close binary scores.

Table 3. R-Precision of the three best runs: binary measure scores and two BDM measure scores.

	GATE-03	NCS01	NCS02
Binary	0.4363	0.4314	0.4314
BDM_high	0.6194	0.6241	0.6241
BDM_low	0.5797	0.5792	0.5792

From Figures 3, 5 and 7 we can easily see the two runs GATE-01 and GATE-02 had much bigger differences between the binary score and the BDM scores than other runs. That is because the two runs adopted a hierarchical classification learning algorithms, while other runs such as GATE-03 and GATE-04 used flat classification. Please refer to [10] for the detailed description of the four runs from the GATE group.

By using the hierarchical classification algorithm, if an instance is not classified correctly, the system would classify it into the class which is close to the true class. In contrast the flat classification does not consider the relations between the classes at all. On the other hand, the BDM measure takes into account the exact matches as well as the partial matches while the binary measure only takes into account the exact matches. Since the hierarchical classification had many more partial matches between two close classes than the flat classification, the former resulted in much higher difference between the BDM score and binary score than the latter. However, unfortunately, because the algorithm used in the two runs GATE-01 and GATE-02 resulted in so much fewer exact matches than other runs such as GATE-03 and GATE-04 (refer to [10] for an analysis of the reasons), their BDM scores were not as high as those of the top runs, although they increased much more than the scores of other runs.

4 Conclusions

We extend the three important evaluation measures, A-Precision, R-Precision and F-measure for document classification and information retrieval to make them take into account the relations between classes in the case of hierarchical classification tasks. We also adapt the generalised measures to the NTCIR-6 F-term patent classification subtask by using the cost function based on the state-of-the-art BDM measure.

We apply those generalised measures to all runs submitted to the NTCIR-6 patent classification subtask. The results show that the BDM scores did change

the ranking of some of the submitted runs, as assigned by the binary scores, although in many cases the results remained unchanged. One interesting finding is that the BDM score swapped the run with the best binary R-Precision score with the second and third best runs in the ranking. The different classification algorithms adopted by the runs from the GATE group, namely the hierarchical classification vs. the flat classification, were nicely reflected in the differences between the BDM scores and the binary scores of these runs.

The question of which measure (cost-based or binary one) is better is really dependent upon the application in which the measure will be used. If an application is not concerned with scoring partial matches, then the binary measure is more appropriate. On the other hand, if an application needs to take into account the misclassification cost in some form or another, then a cost-sensitive measure would surely be more suitable than the binary one.

5 Acknowledgements

Thanks the organisers of the Workshop for the helpful suggestions on improving the paper. This work is supported by the EU-funded Musing (IST-2004-027097) and KnowledgeWeb (IST-2004-507482) projects.

References

[1] E. Agirre and G. Rigau. Word sense disambiguation using conceptual density. In *Proc. of 16th International Conference on Computational Linguistics*, volume 1, pages 16–23, Copenhagen, Denmark, 1996.

[2] C. Brewster, H. Alani, S. Dasmahapatra, and Y. Wilks. Data Driven Ontology Evaluation. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2004)*, Lisbon, Portugal, 2004.

[3] N. Cesa-Bianchi, C. Gentile, A. Tironi, and L. Zaniboni. Incremental Algorithms for Hierarchical Classification. In *Neural Information Processing Systems*, 2004.

[4] P. Cimiano, S.Staab, and J. Tane. Automatic Acquisition of Taxonomies from Text: FCA meets NLP. In *Proceedings of the ECML/PKDD Workshop on Adaptive Text Extraction and Mining*, pages 10–17, Cavtat-Dubrovnik, Croatia, 2003.

[5] J. Davies, D. Fensel, and F. van Harmelen, editors. *Towards the Semantic Web: Ontology-driven Knowledge Management*. Wiley, 2002.

[6] O. Dekel, J. Keshet, and Y. Singer. Large Margin Hierarchical Classification. In *Proceedings of the 21st International Conference on Machine Learning (ICML-2004)*, Canada, 2004.

[7] J. Euzenat. Evaluating ontology alignment methods. In Y. Kalfoglou, M. Schorlemmer, A. Sheth, S. Staab, and M. Uschold, editors, *Semantic Interoperability and Integration*, number 04391 in Dagstuhl Seminar Proceedings, 2005.

[8] U. Hahn and K. Schnattinger. Towards text knowledge engineering. In *Proc. of 15th National Conference on Artificial Intelligence (AAAI-98)*, pages 524–531, Menlo Park, CA, 1998. MIT Press.

[9] M. Iwayama, A. Fujii, and N. Kando. Overview of Classification Subtask at NTCIR-5 Patent Retrieval Task. In *Proceedings of NTCIR-5 Workshop Meeting*, 2005.

[10] Y. Li, K. Bontcheva, and H. Cunningham. SVM Based Learning System for F-term Patent Classification. In *Proceedings of the Sixth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*, 2007.

[11] D. Maynard. Benchmarking ontology-based annotation tools for the semantic web. In *UK e-Science Programme All Hands Meeting (AHM2005) Workshop "Text Mining, e-Research and Grid-enabled Language Technology"*, Nottingham, UK, 2005.

[12] D. Maynard, W. Peters, and Y. Li. Metrics for evaluation of ontology-based information extraction. In *WWW 2006 Workshop on "Evaluation of Ontologies for the Web" (EON)*, Edinburgh, Scotland, 2006.

[13] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support Vector Machine Learning for Interdependent and Structured Output Spaces. In *Proceedings of the 21 st International Conference on Machine Learning*, Banff, Canada, 2004.

[14] E. M. Voorhees. Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. In *Proceedings of the 21st annual international ACM SIGIR*, 1998.

[15] Wessa. Kendall tau Rank Correlation (v1.0.7) in Free Statistics Software (v1.1.21). http://www.wessa.net/rwasp_kendall.wasp/, 2007.

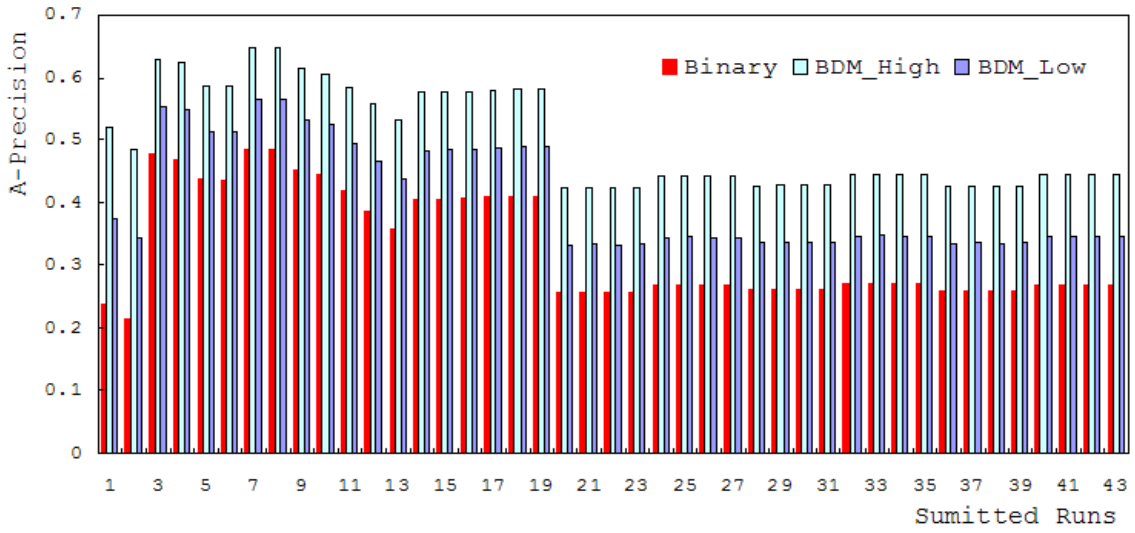

Figure 2. The A-precision evaluation results of all 43 submitted runs for the F-term patent classification subtask at NTCIR-6 by using the cost-based measures *BDM_High* and *BDM_Low*, together with the results using the official evaluation script which uses binary cost. The horizontal axis is for the following systems (from right to left): GATE01, GATE02, GATE03, GATE04, JSPAT01, JSPAT02, NCS01 ,NCS02, NICT01, NICT02 ,NICT03 ,NICT04 ,NICT05 ,NUT01, NUT02, NUT03, NUT04, NUT05, NUT06, RDNDC01, RDNDC02, RDNDC03, RDNDC04, RDNDC05, and RDNDC06, and RDNDC07–24.

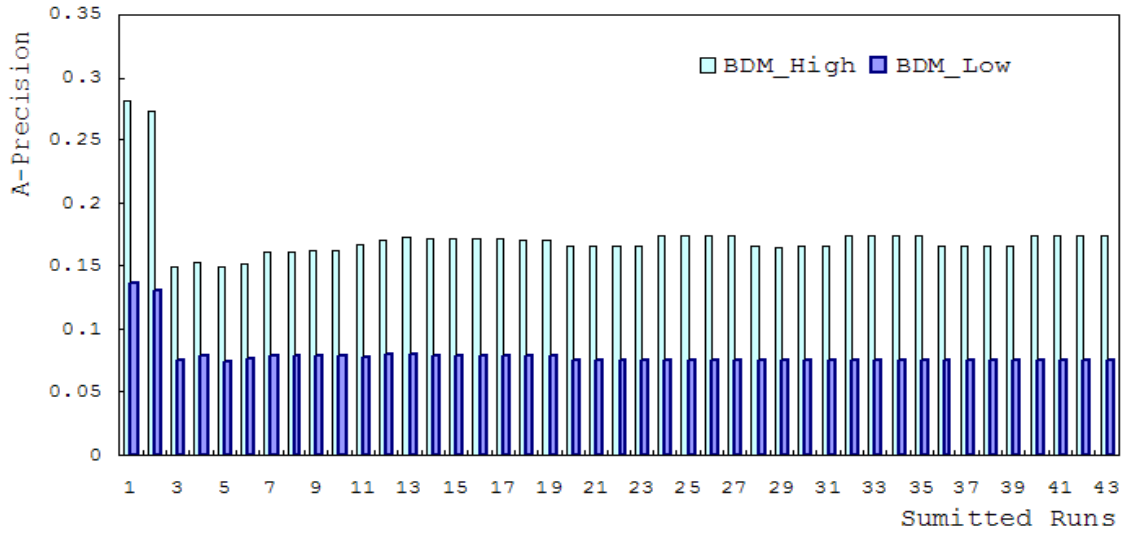

Figure 3. The differences of the A-precision evaluation results between the BDM based scores *BDM_High* (and *BDM_Low*) and the binary based scores for all the 43 submitted runs for the NTCIR-6 F-term patent classification subtask. The horizontal axis is for the 43 runs which are the same as those in Figure 2.

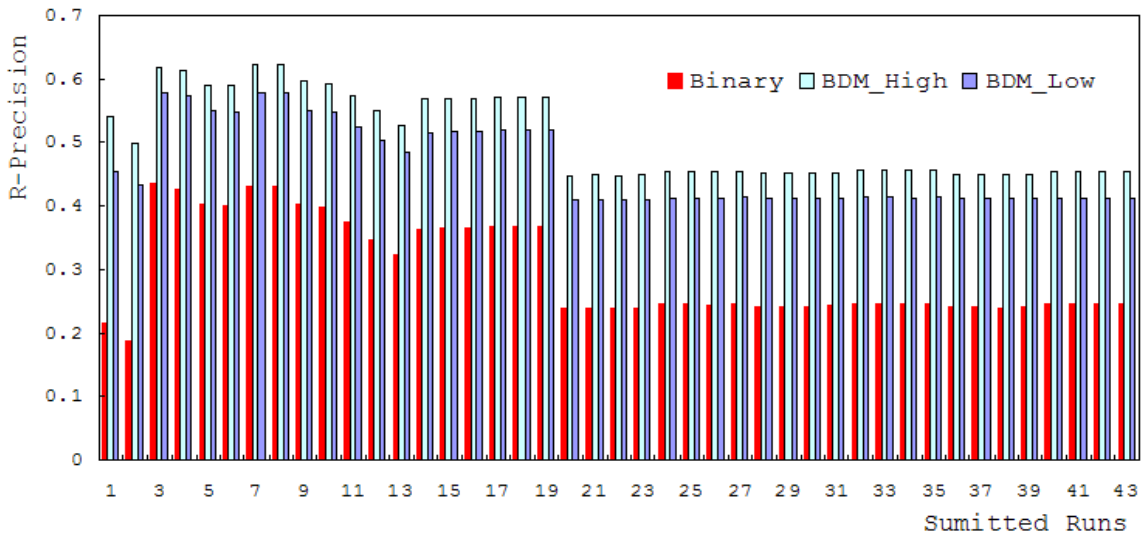

Figure 4. The R-Precision evaluation results of all the 43 submitted runs for the F-term patent classification subtask at NTCIR-6 by using the cost based measures *BDM_High* and *BDM_Low*, together with the results using the official evaluation script which using the binary cost. The horizontal axis is for the 43 runs which are the same as those in Figure 2.

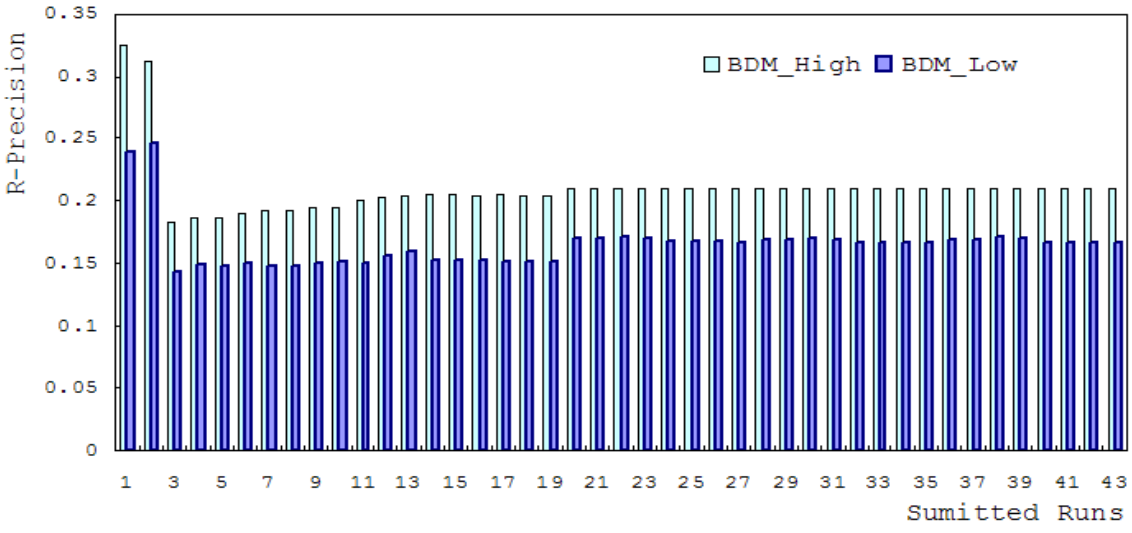

Figure 5. The differences of the R-precision evaluation results between the BDM based scores *BDM_High* (and *BDM_Low*) and the binary based measures for all the 43 submitted runs for the NTCIR-6 F-term patent classification subtask. The horizontal axis is for the 43 runs which are the same as those in Figure 2.

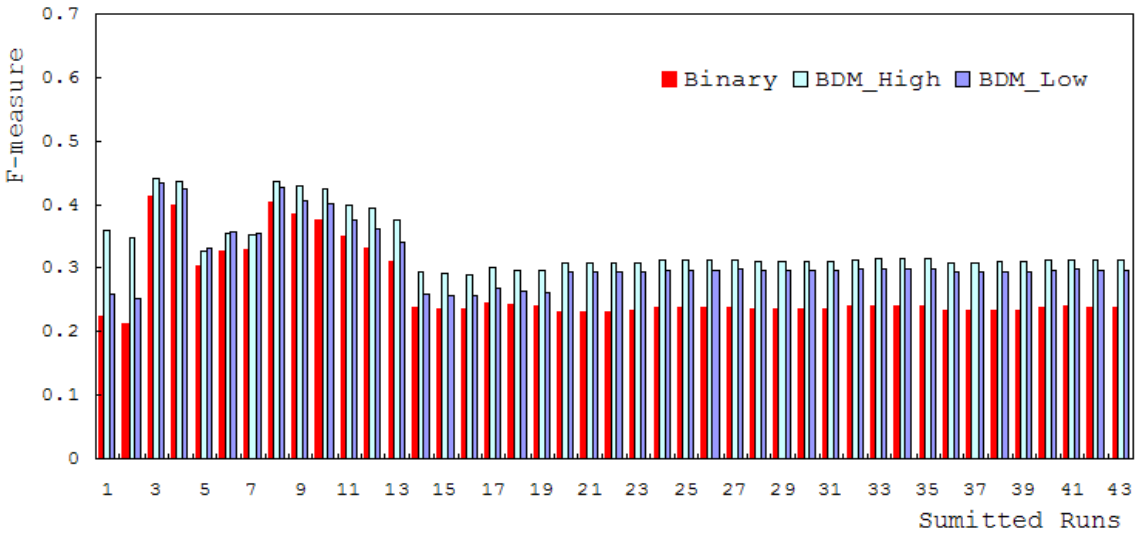

Figure 6. The F-measure evaluation results of all the 43 submitted runs for the F-term patent classification subtask at NTCIR-6 by using the cost based measures *BDM_High* and *BDM_Low*, together with the results using the official evaluation script which using the binary cost. The horizontal axis is for the 43 runs which are the same as those in Figure 2.

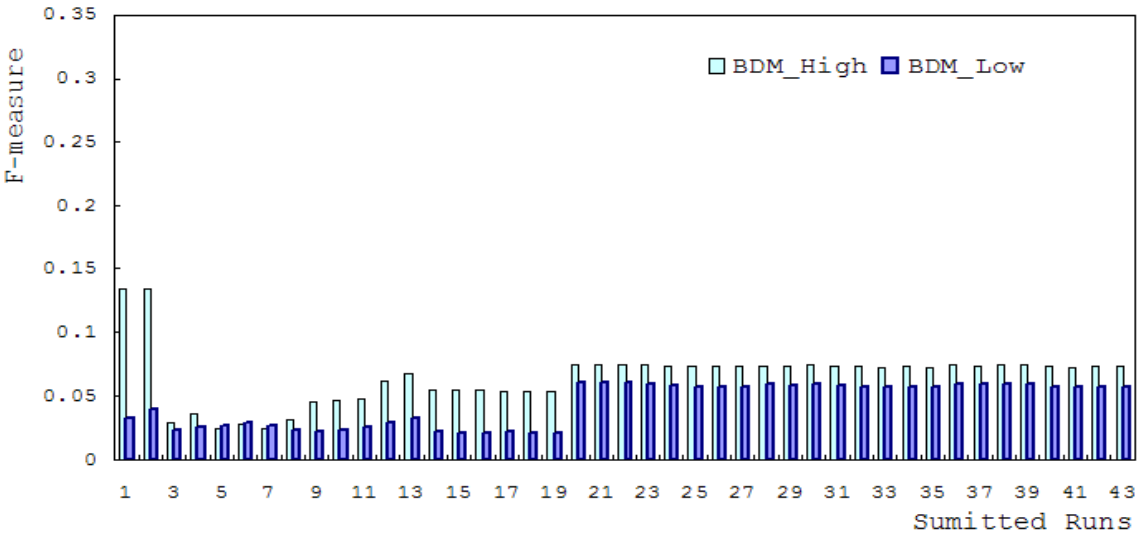

Figure 7. The differences of the F-measure evaluation results between the BDM based scores *BDM_High* (and *BDM_Low*) and the binary based scores for all the 43 submitted runs for the NTCIR-6 F-term patent classification subtask. The horizontal axis is for the 43 runs which are the same as those in Figure 2.