**Table 8: Selected examples of miss-classified paper titles.**

| Paper Title | Conference | Target Class | Top-1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|---|
| A New Algorithm for Optimal Bin Packing | AAAI | AI | ALG | AI | MOD | COLT | DNA |
| (Im)possibility of Safe Exchange Mechanism Design | AAAI | AI | NET | SC | LG | DB | MD |
| Performance Issues and Error Analysis in an Open-Domain Question Answering System | ACL | LG | AI | LG | ALG | DC | SC |
| Active Learning for Statistical Natural Language Parsing | ACL | LG | AI | LG | NN | COLT | ALG |
| Improving Machine Learning Approaches to Coreference Resolution | ACL | LG | AI | LG | ALG | FM | NN |
| A language modelling approach to relevance profiling for document browsing | JCDL | LIS | AI | UI | LG | LIS | ALG |
| Structuring keyword-based queries for web databases | JCDL | LIS | AI | LIS | DB | ALG | ARC |
| A multilingual, multimodal digital video library system | JCDL | LIS | LG | UI | LIS | ECAD | NET |
| SOS: Secure Overlay Services | SIGCOMM | NET | SC | NET | MC | OS | DC |

AI :Artificial Intelligence
ALG :Algorithms
ARC :Architecture
COLT:Computational Learning Theory
DB :Databases
DC :Distributed Computing
DNA :DNA-Based Computing
ECAD:Electronic Computer Aided Design
FM :Formal Methods
LG :Linguistics
LIS :Library and Information Science
MC :Mobile Computing
MOD:Modeling
NET :Networks
NN :Neural Network
OS :Operating Systems
SC :Security
UI :User Interface

**Table 9: Yahoo!'s Computer Science experiment when the corpus size increases. Approach 1.**

| $N_{max}$ | Text Type | Top-1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| 100 | Full Article | .3389 | .3785 | .6045 | .6214 | .6779 |
| | Short Document | .5780 | .7008 | .8034 | .8146 | .8483 |
| | Text Segment | .4917 | .6346 | .6943 | .7242 | .7545 |
| 200 | Full Article | .5311 | .6271 | .6723 | .6949 | .7118 |
| | Short Document | .5780 | .6678 | .7008 | .7409 | .8034 |
| | Text Segment | .4850 | .6213 | .6910 | .7243 | .7409 |
| 400 | Full Article | .4294 | .5028 | .5593 | .6102 | .6251 |
| | Short Document | .5563 | .6632 | .6803 | .7423 | .8011 |
| | Text Segment | .4518 | .5880 | .6545 | .6910 | .7043 |
| 600 | Full Article | .4294 | .5198 | .5593 | .5819 | .5875 |
| | Short Document | .5454 | .6553 | .6731 | .7004 | .7321 |
| | Text Segment | .4219 | .5747 | .6445 | .6678 | .6810 |
| 800 | Full Article | .4294 | .5198 | .5593 | .5819 | .5875 |
| | Short Document | .5450 | .6345 | .6855 | .6921 | .6999 |
| | Text Segment | .4219 | .5083 | .5648 | .6047 | .6146 |

be used to create more value-added Web information services. For common human users, *LiveClassifier* also bestows much convenience. No longer troubled by the tedious work of preparing corpora, users may effortlessly construct many classifiers by his/her own preference.

The effectiveness of *LiveClassifier* deserves some remarks. As discussed in the preceding section, downloading un-labelled Web corpora to augment features or to enhance the size of training corpora has been tried in many recent works. However, few have considered the problem of "how" to collect and organize the corpora.

One may entertain the idea that **HCQF** simply depends on the enormous size of Web resource to train the topic-hierarchy, however, this is not the case. Table 9 lists the results of the Computer Science experiment when training corpora increased. It can be observed that the performance *did not* ameliorate with the size of the training corpora, on the contrary, it is the other way around.

A probable reason of this phenomenon is that the lowly-ranked snippets contain much more noise, thus dragging down the performance. Obviously, downloading Web documents indiscriminately does not ensure success in training. The reason that **HCQF** can get better results is rather its exploiting structural information contained in topic hierarchies. We have presented in Section 3 that subtrees of limited depth extracted from Yahoo!'s directory can achieve satisfactory results. We have also proven in Section 3.2 that in different granularities and in diverse domains, **HCQF** can achieve acceptable results. However, designing experiments of larger scale is still desirable.

*LiveClassifier* can be accessed online in the following URL http://liveclassifier.iis.sinica.edu.tw/. Users can create and modify classifiers online.

# 6. REFERENCES


[1] E. Agirre, O. Ansa, E. Hovy, and D. Martinez. Enriching very large ontologies using the www. In *Proceedings of ECAI 2000 Workshop on Ontology Learning*, 2000.

[2] Z. Bar-Yossef and S. Rajagopalan. Template detection via data mining and its applications. In *Proceedings of the 11st International World Wide Web Conference*, pages 26–33, 2002.

[3] C. Carpineto, R. De Mori, G. Romano, and B. Bigi. An information-theoretic approach to automatic query expansion. *ACM Transactions on Information Systems*, 19(1):1–27, 2001.

[4] C.-H. Chang and S.-L. Lui. Iepad: information extraction based on pattern discovery. In *Proceedings of 10th International World Wide Web Conference*, pages 681–688, 2001.

[5] S.-L. Chuang and L.-F. Chien. Towards automatic generation of query taxonomy: A hierarchical query clustering approach. In *Proceedings of the 2nd IEEE International Conference on Data Mining*, pages 75–82, 2002.

[6] W. Cohen and W. Fan. Learning page-independent heuristics for extracting data from web pages. In *Proceedings of the 8th International World Wide Web Conference*, 1999.

[7] W. Cohen, M. Hurst, and L. Jensen. A flexible learning system for wrapping tables and lists in html documents. In *Proceedings of the 11th International World Wide Web Conference*, pages 232–241, 2002.

[8] W. Cohen and Y. Singer. Context-sensitive learning methods for text categorization. In H.-P. Frei, D. Harman, P. Schäuble, and R. Wilkinson, editors, *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 307–315, Zürich, CH, 1996. ACM Press, New York, US.

[9] B. V. Dasarathy. *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. McGraw-Hill Computer Science. IEEE Computer Society Press, Las Alamitos, California, 1991.

[10] E. O. Doorenbos, R. and D. Weld. A scalable comparison-shopping agent for the world-wide web. In *Proceedings of Autonomous Agents*, pages 39–48, 1997.

[11] C.-N. Hsu and M.-T. Dung. Generating finite-state transducers for semi-structured data extraction from the web.