# KDD CUP-2005 Report: Facing a Great Challenge

Ying Li
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052
yingli@microsoft.com

Zijian Zheng[1]
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052
zijian@acm.org

Honghua (Kathy) Dai
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052
kathydai@microsoft.com

## ABSTRACT

The KDD-Cup 2005 Competition was held in conjunction with the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. The task of the KDD-Cup 2005 competition was to classify 800,000 internet user search queries into 67 predefined categories. This task is easy to understand, but the lack of straightforward training set, subjective user intents of queries, poor information in short queries, and high noise level make the task very challenge.

In this paper, we summarize the competition task, the evaluation method, and the results of the competition. Here we only highlight some key techniques used in submitted solutions. The technical details of the solutions from the three award winning teams are available in their papers separately in this issue of SIGKDD Explorations. At the end, we also share the results of a survey conducted with this year's Cup participants. To facilitate research in this area, the task description, data, answer set, and related information of this KDD-Cup are published at the KDD-Cup 2005 web site: http://www.acm.org/sigs/sigkdd/kdd2005/kddcup.html.

## Keywords

KDD-Cup, search, query, categorization, user intent, machine learning, data mining, text classification, text mining, knowledge discovery

## 1. INTRODUCTION

Computer search including internet search has become a part of many people's daily life and work. Given the exponential growth of information's availability in electronic form, search becomes one of the most important and effective approaches to finding correct/relevant information to serve our needs. A user can type in key words in a search engine to find out where to buy a product and whether a certain price is a good price. A user can also find travel attractions to fit his/her interests. If a user is interested in certain medicine, he/she can find plenty related information including its usage and potential side-effects. Researchers can easily find the latest development of a research topic. To plan a hiking trip for a coming weekend, one can find the weather forecast simply through a search. These are just a few examples of how search can help a user's daily life.

Although researchers and industry practitioners have achieved tremendous success in developing smart search engines, we are still facing many great challenges as current search engines are not very accurate. The difference is still quite big between what search engines can do and what we expect them to do. It is not uncommon that a search engine returns irrelevant, misleading or, incorrect results after you type in a query. In another time, the relevance results are returned but down to the bottom of a long result list.

Due to the nature that huge amount data is available from each search engine and many problems of search can be turned into learning or modeling problems, there is a great potential for data mining techniques to contribute to the success of search.

Since late 90's, researchers and practitioners have been studying search query data [8], trying to find search patterns [8], understanding search user intents [10][11][12], and providing personalized search [13][14][15][16]. A survey on search related research is available [9].

The other side of search being a difficult problem is that the information contained inside the data is often incomplete, fuzzy, and indirect. The intents of search engine users are highly subjective. They are sometimes ambiguous and dynamic. All of these present big challenges to the data mining community. In KDD-Cup 2005 [2], we presented one challenge problem: Search query categorization.

Text classification and categorization is a well-known topic in Information Retrieval and Text Mining fields. Manning and Schtüze [5] discusses general methodologies and applications of text categorization. Most work in this area has been focused on categorizing Web pages or longer text or corpus. However, search query classification is very different in the sense that queries are usually very short on the one hand, and with implicit and subjective user intents on the other hand. Therefore, how to automatically understand user search intents given the search queries would be very interesting to IR and text mining researchers.

In this report, we first describe the competition task presented to the participants in Section 2, including a discussion on why this task is challenging. In Section 3, we present the evaluation method. Then, we highlight the interesting techniques from the submissions in Section 4, and analyze the overall results as a whole in Section 5. The detailed presentations of techniques from the three winning teams are available as three separate papers in this issue of SIGKDD explorations. Finally, we summarize in Section 6.

## 2. THE TASK

This year's competition is in the area of search query categorization. The task was to classify 800,000 search queries into 67 predefined categories. We provided:

[1] Zijian Zheng was at Amazon.com during the time of KDD-Cup 2005.

- 67 categories,
- 800,000 search query strings, and
- A very small set of 111 queries with labeled categories.

Participants are required to tag each of the 800,000 queries with up to 5 categories. This task description and dataset are available at KDD-Cup 2005 web site: **http://www.acm.org/sigs/sigkdd/kdd2005/kddcup.html.**

The 67 categories are listed in Table 1. They are hierarchical with two levels. The top level has 7 categories. Each of the top level has several second level sub categories. They cover most of the areas in the internet information space. The purpose here is to provide a reasonable set of categories to make the competition meaningful in real life. We do not claim that this is the best category set in any sense.

**Table 1. The 67 predefined categories**

| |
|---|
| Computers\Hardware |
| Computers\Internet & Intranet |
| Computers\Mobile Computing |
| Computers\Multimedia |
| Computers\Networks & Telecommunication |
| Computers\Security |
| Computers\Software |
| Computers\Other |
| Entertainment\Celebrities |
| Entertainment\Games & Toys |
| Entertainment\Humor & Fun |
| Entertainment\Movies |
| Entertainment\Music |
| Entertainment\Pictures & Photos |
| Entertainment\Radio |
| Entertainment\TV |
| Entertainment\Other |
| Information\Arts & Humanities |
| Information\Companies & Industries |
| Information\Science & Technology |
| Information\Education |
| Information\Law & Politics |
| Information\Local & Regional |
| Information\References & Libraries |
| Information\Other |
| Living\Book & Magazine |
| Living\Car & Garage |
| Living\Career & Jobs |
| Living\Dating & Relationships |
| Living\Family & Kids |
| Living\Fashion & Apparel |
| Living\Finance & Investment |
| Living\Food & Cooking |
| Living\Furnishing & Houseware |
| Living\Gifts & Collectables |
| Living\Health & Fitness |
| Living\Landscaping & Gardening |
| Living\Pets & Animals |
| Living\Real Estate |
| Living\Religion & Belief |
| Living\Tools & Hardware |
| Living\Travel & Vacation |
| Living\Other |
| Online Community\Chat & Instant Messaging |
| Online Community\Forums & Groups |
| Online Community\Homepages |
| Online Community\People Search |
| Online Community\Personal Services |
| Online Community\Other |
| Shopping\Auctions & Bids |
| Shopping\Stores & Products |
| Shopping\Buying Guides & Researching |
| Shopping\Lease & Rent |
| Shopping\Bargains & Discounts |
| Shopping\Other |
| Sports\American Football |
| Sports\Auto Racing |
| Sports\Baseball |
| Sports\Basketball |
| Sports\Hockey |
| Sports\News & Scores |
| Sports\Schedules & Tickets |
| Sports\Soccer |
| Sports\Tennis |
| Sports\Olympic Games |
| Sports\Outdoor Recreations |
| Sports\Other |

The 800,000 search queries were randomly selected from MSN [3] search logs with some preliminary filtering.

The small set of 111 queries with tagged categories (up to five) was provided as a sample to demonstrate the mapping between query strings to categories. It is not intended to be used as a training set in the conventional training/test sense. Of course, participants are free to use the information in the set in any way they like.

The goal of this year's KDD-Cup is to challenge the data mining community with this open resource problem. Although every one who used search engine can understand the task, it presents the following six challenges:

- **No straightforward training data**: Not like previous years' KDD-Cups, where for most tasks standard learning algorithms can be directly applied, people cannot directly apply standard learning algorithms to this year's task since no training data is provided. The tiny set of labeled queries is provided only for showing the semantics of the categories.
- **Open resource**: Although the resources and information provided with the task are limited, there are many other resources where participants can acquire rich related information for solving the problem. For example, the queries are usually very short comparing to text documents. In order to understand the context or semantics of the queries, participants may need to solicit extra information from various media, such as search engine, dictionary, thesaurus, and document repository.
- **Subjective search query intents**: The meanings of search queries or intents of search users are subjective. For example, the query "Saturn" might mean the Saturn car to some users and mean Saturn, the planet to others. The meanings of a query may also be different for the same person at a different time.

- **Implicit category semantics**: Since we didn't provide detailed description on each category, participant have to reply on common sense to acquire the implicit category semantics.
- **Noise in data**: The provided query data was from real search query logs. It contains a lot of human introduced noises such as misspelling and incorrect word breaking.
- **Scalability and automation**: It is very time consuming and costly to manually categorize the queries. A scalable automatic solution is required to label the 800,000 queries in two months time.

To solve this problem, participants must develop/use multiple technologie and acquire extra information externally. They are opened to all the resources they can access. This simulates the real life search engine (or any related components) development well.

# 3. EVALUTAION CRETERIA AND METHOD

In this year's KDD-Cup, we used three evaluation criteria on each solution submitted by participants: Precision, F1, and creativity. Precision and F1 are defined below. They are commonly used measures in the area like information retrial and data mining [4][5]. Creativity was judged by reviewers based on creative ideas in the solution, its scalability, and the level of automation.

$$\text{Precision} = \frac{\sum_i \#\ of\ queries\ are\ correctly\ tagged\ as\ c_i}{\sum_i \#\ of\ queries\ are\ tagged\ as\ c_i}$$

$$\text{Recall} = \frac{\sum_i \#\ of\ queries\ are\ correctly\ tagged\ as\ c_i}{\sum_i \#\ of\ queries\ whose\ category\ is\ labeled\ as\ c_i\ by\ human\ labeler}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Three awards were set with one for each of the three criteria. They are:

- Query Categorization Precision Award,
- Query Categorization Performance Award, and
- Query Categorization Creativity Award

In the remaining part of this section, we describe the submission requirements for cup participants, submission evaluation process, and award winner selection process.

### *3.1.1 Submission requirements*

There were about 70 days between the task and dataset available online and the solution submission deadline. The required submission from each participant team consists of:

- Category tags for the 800,000 queries with up to five categories for each query (the order of the tagged categories for each query is not important) and
- A description of the techniques/algorithms used for categorizing the queries (up to five pages in length).

Due to the fact that a system can be tuned differently for optimizing precision and F1 respectively, we allowed each team to submit one solution for the Precision Award and one solution for the Performance Award. In this case, the solution dedicated for the Precision Award is not qualified for the Performance Award even its F1 score is high. Similarly the solution dedicated for the Performance Award is not qualified for the Precision Award even its precision is high.

Some teams raised an issue that they do not want to share their technique details since those are their companies' confidential intellectual property. This is understandable, and we gave participants the option of not submitting solution descriptions. Any team that did not submit solution descriptions would not be considered for the Creativity Award.

### *3.1.2 Evaluation Process*

To evaluate a solution submitted from a participant, we need to know the correct or standard category labels of the 800,000 queries. One way to get this data is through human labeling. However, manually labeling all the 800,000 queries is too expensive and time consuming. The approach we took was:

- Randomly select 800 evaluation queries from the 800,000 queries. In the random selection process, if a junk query (e.g. a string without meaning) or a non-English query was selected, another query was randomly selected until all the 800 queries are non-junk English queries. This process resulted in an evaluation query set of 800 queries.
- Ask three human editors to label the entire evaluation query set manually using the 67 categories separately. This generated three answer sets each containing the 800 evaluation queries with up to five categories. Due to the fact that the human labeling process is subjective and each person's knowledge is incomplete, we used three human editors to reduce the variance of evaluation scores, and make the evaluation reasonably fair.
- For each submitted solution, we found the 800 evaluation queries with the participant's category tags. We compared these category tags against the three answer sets to compute three precision scores and three F1 scores. We then used the following formulas to compute an overall precision and an over all F1 score.

$$\text{Overall Precision} = \frac{1}{3}\sum_{i=1}^{3}(\text{Precision against Answer\_Set\_i})$$

$$\text{Overall F1} = \frac{1}{3}\sum_{i=1}^{3}(\text{F1 against Answer\_Set\_i})$$

The participants were made aware of this evaluation process together with the task description, but they did not know which subset of the 800,000 queries will be used in the final evaluation. After the competition, we published the answer sets at the KDD-Cup 2005 web site: http://www.acm.org/sigs/sigkdd/kdd2005/kddcup.html

Following this evaluation process, we obtained the overall precision and F1 score values of all submitted solutions. They are listed in Table 2. Each submitted solution was given a unique ID. These evaluation results are used to select award winning teams.

For the Creativity Award, a usual paper review approach was used. We had four reviewers. Each of them reviewed the submitted technology/algorithm descriptions of all teams separately. The review focused on the creative ideas used for solving the competition task, the scalability of the solutions, and how easy the solution can be automated.

**Table 2. Overall precision and F1 score values of all submitted solutions**

| ID | Precision | F1 | ID | Precision | F1 |
|---|---|---|---|---|---|
| 1 | 0.145099 | 0.146839 | 20 | 0.127784 | 0.126848 |
| 2 | 0.116583 | 0.139732 | 21 | 0.340883 | 0.340090 |
| 3 | 0.339435 | 0.309754 | 22 | 0.414067 | 0.444395 |
| 4 | 0.110885 | 0.124228 | 23 | 0.237661 | 0.250293 |
| 5 | 0.310680 | 0.085639 | 24 | 0.244565 | 0.258035 |
| 6 | 0.254815 | 0.246264 | 25 | 0.753659 | 0.205391 |
| 7 | 0.263953 | 0.306359 | 26 | 0.255726 | 0.274579 |
| 8 | 0.454068 | 0.405453 | 27 | 0.206919 | 0.205302 |
| 9 | 0.264312 | 0.306612 | 28 | 0.148503 | 0.176140 |
| 10 | 0.334048 | 0.342248 | 29 | 0.171081 | 0.198500 |
| 11 | 0.107045 | 0.116521 | 30 | 0.145467 | 0.173173 |
| 12 | 0.196117 | 0.207787 | 31 | 0.108305 | 0.108174 |
| 13 | 0.326408 | 0.357127 | 32 | 0.169620 | 0.232654 |
| 14 | 0.317308 | 0.312812 | 33 | 0.469353 | 0.255096 |
| 15 | 0.271791 | 0.265450 | 34 | 0.198284 | 0.191618 |
| 16 | 0.050918 | 0.060285 | 35 | 0.320750 | 0.384136 |
| 17 | 0.264009 | 0.218436 | 36 | 0.211284 | 0.129937 |
| 18 | 0.206167 | 0.247854 | 37 | 0.423741 | 0.426123 |
| 19 | 0.136541 | 0.127008 | | | |

### *3.1.3 Award Winner Selection Process*

For each of the three evaluation criteria, two teams were selected to receive the corresponding award and runner-up respectively based on their evaluation results on the criterion.

- Query Categorization Performance Award

All the submitted solutions dedicated for this award by the participating teams were qualified. A single solution from a team was qualified by default. We ranked all qualified solutions based on their overall F1 scores as shown in Table 3. The two teams of the top two solutions were selected as the winners of this award.

- Query Categorization Precision Award

Similar to the previous award, all the submitted solutions dedicated for this award by the participating teams were qualified. A single solution from a team was qualified by default. We first ranked all these qualified solutions using overall F1 scores. The top 10 solutions became the candidates. Then, we ranked these 10 solutions using the overall Precision values as shown in Table 4. The two teams of the top 2 solutions in this ranked list were selected as the winners of this ward.

**Table 3. Qualified solutions for the Performance Award ranked on overall F1**

| ID | F1 | Precision | ID | F1 | Precision |
|---|---|---|---|---|---|
| 22 | 0.444395 | 0.414067 | 17 | 0.218436 | 0.264009 |
| 8 | 0.405453 | 0.454068 | 12 | 0.207787 | 0.196117 |
| 35 | 0.384136 | 0.320750 | 27 | 0.205302 | 0.206919 |
| 13 | 0.357127 | 0.326408 | 29 | 0.198500 | 0.171081 |
| 10 | 0.342248 | 0.334048 | 34 | 0.191618 | 0.198284 |
| 21 | 0.340090 | 0.340883 | 28 | 0.176140 | 0.148503 |
| 14 | 0.312812 | 0.317308 | 30 | 0.173173 | 0.145467 |
| 9 | 0.306612 | 0.264312 | 1 | 0.146839 | 0.145099 |
| 7 | 0.306359 | 0.263953 | 2 | 0.139732 | 0.116583 |
| 26 | 0.274579 | 0.255726 | 36 | 0.129937 | 0.211284 |
| 24 | 0.258035 | 0.244565 | 19 | 0.127008 | 0.136541 |
| 33 | 0.255096 | 0.469353 | 20 | 0.126848 | 0.127784 |
| 23 | 0.250293 | 0.237661 | 4 | 0.124228 | 0.110885 |
| 18 | 0.247854 | 0.206167 | 11 | 0.116521 | 0.107045 |
| 6 | 0.246264 | 0.254815 | 31 | 0.108174 | 0.108305 |
| 32 | 0.232654 | 0.169620 | 16 | 0.060285 | 0.050918 |

**Table 4. Candidate solutions for the Precision Award ranked on overall Precision**

| ID | Precision | F1 | ID | Precision | F1 |
|---|---|---|---|---|---|
| 37 | 0.423741 | 0.426123 | 35 | 0.320750 | 0.384136 |
| 21 | 0.340883 | 0.340090 | 15 | 0.271791 | 0.265450 |
| 3 | 0.339435 | 0.309754 | 9 | 0.264312 | 0.306612 |
| 10 | 0.334048 | 0.342248 | 7 | 0.263953 | 0.306359 |
| 13 | 0.326408 | 0.357127 | 26 | 0.255726 | 0.274579 |

- Query Categorization Creativity Award

After finishing the individual reviews on the submitted descriptions, the four reviewers had a meeting to thoroughly discuss which solutions were the best and the second best among the submissions with the overall F1 within the top 20. The two teams of these two solutions became the winners of this award.

## 4. TECHNIQUES FROM SUBMISSIONS

We observed good techniques and interesting ideas from the submitted solution descriptions. They differ mostly in the approaches for designing the solutions and the learning methods used. Here, we highlight the major techniques adopted by most participants.

## 4.1 Common Approaches for Categorization

No solutions submitted are the same. We have seen reasonable differences in the approaches used by the contestants. We, however, abstract the solutions from the contestants as consisting of the following three high level components:

1. Preprocessing,
2. Gathering extra information to augment the queries, and
3. Modeling

### *4.1.1 Preprocessing*

There are several big benefits to preprocessing the queries:

1. Reduce the impact of noise in queries;
2. Reduce the workload on noisy queries;
3. Benefit feature extraction; and
4. Improve accuracy

Although some preliminary filtering was performed by the organizers on the raw queries for removing obvious inappropriate content, the participants still have to handle some noisy and versatile forms of data, such as misspelling, foreign language words, acronyms, etc. It is interesting to see people in Data Mining and Information Retrieval fields to apply existing text mining preprocessing techniques and create new methodologies for processing search queries specifically.

The common methods adopted by contestants are mostly standard text mining techniques: stop words filtering, stemming, and term frequency filtering, etc.

Some participants also applied more advanced techniques, such as, spelling correction, compound word breaking, abbreviation expansion, and named entity detection.

### *4.1.2 Gathering Extra Information*

Search queries can be treated as a semi-natural-language with implicit and subjective user intents. On the other hand, a search query itself often only contains a few features in a very sparse feature space. Many queries only have one term in it. Sometimes it is hard to learn the meaning of a query solely based on the query itself. Many participants came up with different ways of gathering extra information to augment the query terms.

Participants mostly utilize the unlimited knowledge base available on the World Wide Web and search engines. They used the following resources to build the knowledge base: search engines/Internet search result snippets, bag of words, search engine directories, search result titles, or search result web pages. Some of them also adopted tools like WordNet [6] to expand queries.

Besides gathering extra information for augmenting search queries, some contestants also tried to build semantics for the categories. Tools like WordNet [6] and WikiPedia [7] were used to build category descriptions.

### *4.1.3 Modeling*

This year's competition task provided three major dimensions: search queries, words or phrases, and categories. Participants used the following two major modeling approaches to building the models based on these dimensions.

1. Mapping pre-defined/exisiting directory structure to KDD Cup categories.
2. Constructing the mappings between KDD categories and words (descriptions), and then using the mappings to answer the categories of search queries that were treated as a bag of words.

## 4.2 Frequently Used Learning Methods

Most of the participants adopted machine learning algorithms in their solutions. Among them, Naïve Bayesian classifier, SVM, KNN, Neural Network, Logistic Regression are the popular ones.

A few participants constructed multiple models and then combined them together to achieve better results. Some of them used distance/probability as criteria to combine predictions. Some applied ad hoc rules in combining predictions. Others adopted ensemble learning (e.g. Boosting).

More advanced methods were adopted to tune model parameters, such as using manually tagged examples to tune model parameters or using reward/penalty factors in model tuning/training.

Other learning approaches taken by participants are transforming multi-class problem into multiple binary-class problems and Iterative learning.

## 4.3 Other Interesting Techniques Used

As we have mentioned earlier, search queries can be treated as a semi-natural-language with implicit and subjective intents. Simple preprocessing techniques, such as stemming and stop word filtering, are not sufficient in capturing the meaning of a search query. A lot of query processing techniques were considered by the competition contestants:

- Query / term clustering,
- Detecting centroid (key) words in a query,
- Soundex for query mapping,
- Identifying specific groups for queries (e.g. URL, email, trash, and name entities),
- Extracting information from text appeared in URLs,
- Query structure detection,
- Lexical relational structure in queries, and
- Creating phrase lists.

Some people used substring/partial matching and regular expression matching in text processing and inferences.

Knowledge representation is also an important aspect. The way queries or other knowledge are represented could dramatically affect the classification performance. Here are a few popular representation techniques we saw among submitted solutions:

- N-gram (e.g. bi-gram and tri-gram),
- Feature-value vector representation, and
- Graphic representation (for word relationship in queries).

Performing a systematic study on preprocessing, representation, and mining the semantics and context of search queries will be a very interesting research topic. We hope to see more exciting

research in this field using the data set we contributed to the community.

## 5. RESULT ANALYSIS

At the time KDD-Cup submission deadline, 142 teams registered. Among them, 32 teams submitted 37 solutions. As described in the previous section, a good number of very interesting and creative ideas and techniques were developed and used. As a result of applying these ideas and techniques, some solutions achieved quite impression scores in terms of precision and F1.

In this section, we first present the award winning teams, and then summarize the overall results from all the teams that made submissions. Since we used three labelers in the evaluation process, we will discuss whether the three labelers agree on the award winners. At the end of this section, we will share the results of a survey we conducted with registered participants.

### 5.1 Award Winning Teams

Here are the winning teams and their members of all the awards.

- Query Categorization Precision Award
  - Winner
    Hong Kong University of Science and Technology team (Dou Shen, Rong Pan, Jiantao Sun, Junfeng Pan, Kangheng Wu, Jie Yin and Qiang Yang)
  - Runner-up
    Budapest University of Technology team (Zsolt T. Kardkovács, Domonkos Tikk, Zoltán Bánsághi)
- Query Categorization Performance Award
  - Winner
    Hong Kong University of Science and Technology team (Dou Shen, Rong Pan, Jiantao Sun, Junfeng Pan, Kangheng Wu, Jie Yin and Qiang Yang)
  - Runner-up
    MEDai/AI Insight/ Humboldt University team (David S. Vogel, Steve Bridges, Steffen Bickel, Peter Haider, Rolf Schimpfky, Peter Siemen, Tobias Scheffer)
- Query Categorization Creativity Award
  - Winner
    Hong Kong University of Science and Technology team (Dou Shen, Rong Pan, Jiantao Sun, Junfeng Pan, Kangheng Wu, Jie Yin and Qiang Yang)
  - Runner-up
    Budapest University of Technology team (Zsolt T. Kardkovács, Domonkos Tikk, Zoltán Bánsághi)

It is worth mentioning that the Hong Kong University of Science and Technology team did a great job and won all the three awards.

### 5.2 Result Overview

Among the 37 submitted solutions, the highest F1 is 0.44, the lowest F1 is 0.06, and the median is 0.23. Figure 1 Shows the F1 score distribution. We can see that the F1 scores are nicely normal distributed with a good number of solutions having reasonably high F1 scores.

The range of the 37 precision scores is between 0.75 and 0.05, with a median of 0.24. Figure 2 shows the precision distribution, which is skewed toward low precision scores. It is worth mentioning that one solution has a very high precision score of 0.75. That is a result of heavy turning of the algorithm to make it generate high precision. This solution actually created very few category predictions, one prediction for each query in many cases. However, because its F1 is very low: 0.21 (lower than the median), it was not qualified for the Precision Award.

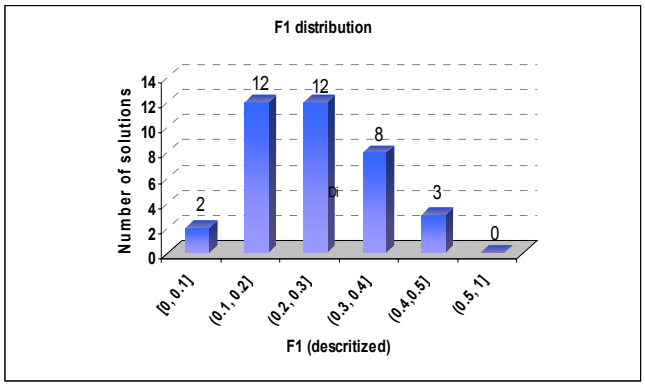


**Figure 1. F1 distribution**

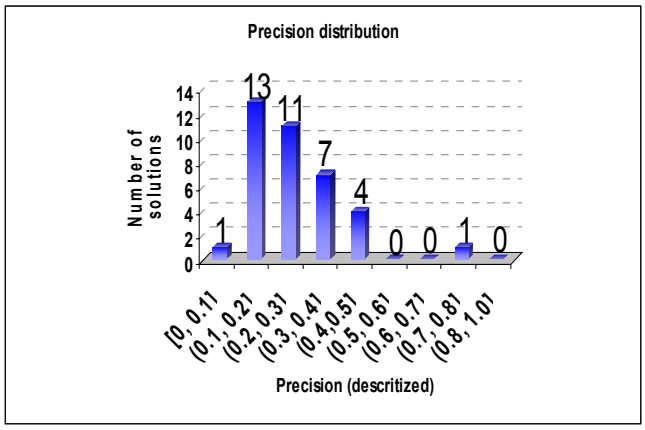


**Figure 2. Precision distribution**

### 5.3 Do Three Labelers Agree on Winners?

As specified before, the three answer sets used in the evaluation process were generated through manual labeling by three human editors separately. Therefore, each answer set reflects the understanding of the problem and the knowledge of the corresponding editor. This naturally raises the following two questions:

1. Do these three labelers agree on the award winners?
2. How much do these three labelers agree with each other?

To answer the first question on the Performance Award, we list the top 10 submissions using their IDs ranked by each labeler as well as the overall score of all the three labelers on F1 in Table 5. We can see that all the three labelers agree on the first three places. They start to disagree at the fourth place. Therefore, the answer is "yes" to this question.

**Table 5. Ranked top 10 F1 score submissions by ID**

| Ranking | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Overall** | **22** | **8** | **35** | **13** | **10** | **21** | **14** | **9** | **7** | **26** |
| Labeler 1 | 22 | 8 | 35 | 13 | 10 | 21 | 14 | 7 | 9 | 26 |
| Labeler 2 | 22 | 8 | 35 | 10 | 21 | 13 | 33 | 9 | 7 | 14 |
| Labeler 3 | 22 | 8 | 35 | 13 | 10 | 21 | 14 | 9 | 7 | 26 |

Similarly, we list the top 10 submissions using their IDs ranked by each labeler as well as the overall score of all the three labelers on precision in Table 6. This time, the three labelers agree at the first place, and start to disagree at the second place. We can see that submission 21 was ranked at the second place by Labeler 2 and at the third place by Labeler 1 and Labeler 3. While submissions 3 and 13 were ranked at the second place by Labeler 1 and Labeler 3 respectively, they were ranked at the fourth place or lower by the other labelers. Therefore, submission 21 should be given the second place although the three labelers had certain disagreement on this.

From this analysis, we can see that the three labelers highly agreed on the winners of the awards although not completely

**Table 6. Ranked top 10 precision submissions by ID**

| Ranking | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Overall** | **37** | **21** | **3** | **10** | **13** | **35** | **15** | **9** | **7** | **26** |
| Labeler 1 | 37 | 3 | 21 | 10 | 13 | 35 | 7 | 9 | 15 | 26 |
| Labeler 2 | 37 | 21 | 10 | 3 | 35 | 15 | 13 | 26 | 9 | 7 |
| Labeler 3 | 37 | 13 | 21 | 3 | 10 | 35 | 15 | 9 | 7 | 26 |

Now, let us look at the question of how much these three labelers agree with each other. Table 7 shows the average precision and F1 score values of each labeler when evaluated against the other two labelers. Both precision and F1 scores of the three labelers are at the level of 0.5, indicating that they have certain disagreement. However, the lowest precision score in this table is higher than the highest precision score from the contestants (Table 4) and the lowest F1 score in this table is higher than the highest F1 score from the contestants (Table 3). This indicates that the agreements among the three labelers are reasonably high for the purpose of judging the submitted solutions.

**Table 7. Average scores of each labeler when evaluated against the other two labelers**

| | Avg Precision | Avg F1 |
|---|---|---|
| **Labeler 1** | 0.500852 | 0.537741 |
| **Labeler 2** | 0.612650 | 0.477077 |
| **Labeler 3** | 0.463077 | 0.512238 |

## 5.4 Participant Survey

We performed a survey at the end of the competition to collect information about the participants and their works. The purpose was to gain more understanding on this year's KDD-Cup. We sent emails with a set of questions to all the registered participants. 30 of them responded, including teams who submitted solutions and teams who did not submit solutions. Since the sample size 30 is relatively small comparing to 142, please read the statistics with caution.

As mentioned earlier, only 32 out of 142 teams submitted solutions. It is interesting to know why so many teams did not submit. From the survey, we learned that the major reason for not submitting was that the teams did not have enough time.

Another interesting question is which countries/regions the participants were from. Figure 3 shows the participants' country/region distribution (unit is team). The responded participants covered 9 countries/regions. Among them, United States, China, and Germany were the top three.

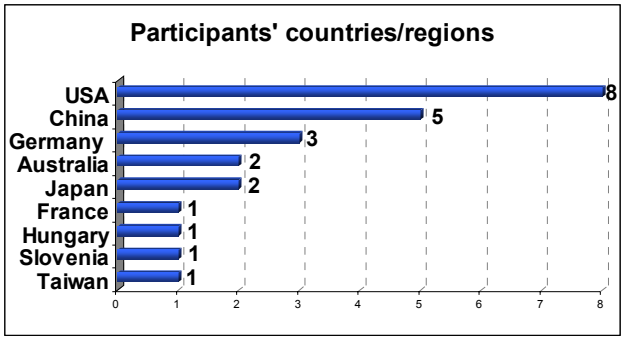


**Figure 3. Participants' county/region distribution (unit: team)**

We knew that the KDD community is a good mixture of people from academy and industry. Figure 4 indicates a similar mixture for this year's KDD-Cup. Although 52% teams were from universities and 39% teams were from companies, the 9% teams marked as individual were likely belong to industry. Therefore, the academy participant teams and industry participant teams were well balanced. We also found from the survey that students and industry professionals were the primary contributors (the team member spent most time in the team on the KDD-Cup task). As shown in Figure 5, each of them was 41%.

From the survey responses, we found that while most common team size was 6 to 10 people (36%), we had significant number of single person teams (27%). The team size distribution is shown in Figure 6.

Due to the great challenge of this KDD-Cup task, many team spent tremendous effort. As Figure 7 shows, 40% teams spent more than 100 person-hours. The person-hour range for all the survey response teams was from 20 to 1000 with the median as 179.25.

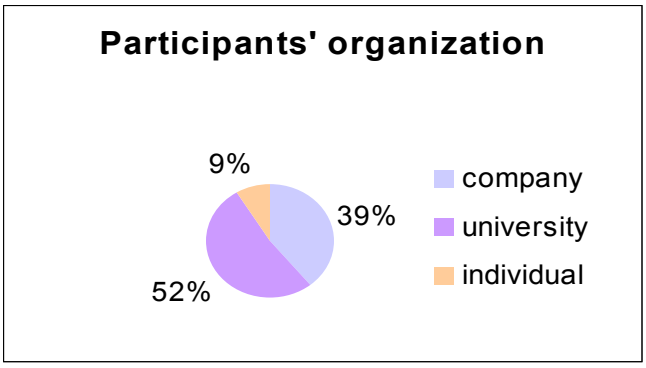


**Figure 4. Organization types of the KDD-Cup participating teams**

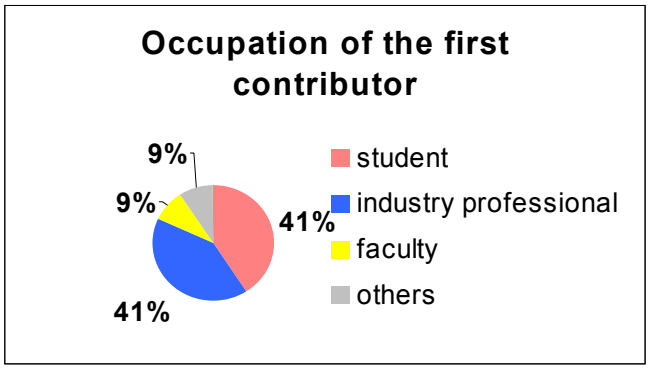


**Figure 5. Occupations of the primary contributors in participating teams**

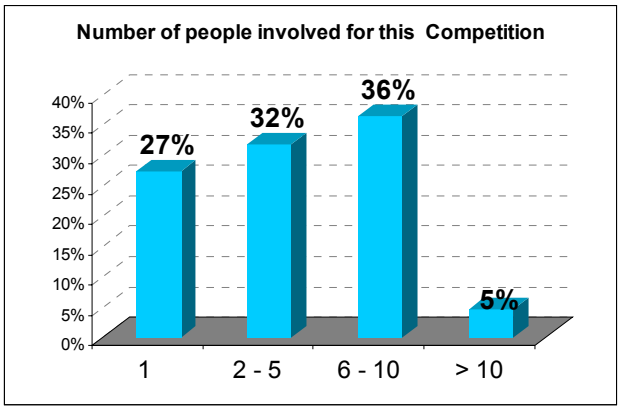


**Figure 6. Team size of the participating teams**

It is known that some people participated in several KDD-Cups. For this year, we can see many return participant teams as in Figure 8. However, the majority survey response teams are first time participants (56%), indicating a healthy dynamic community.

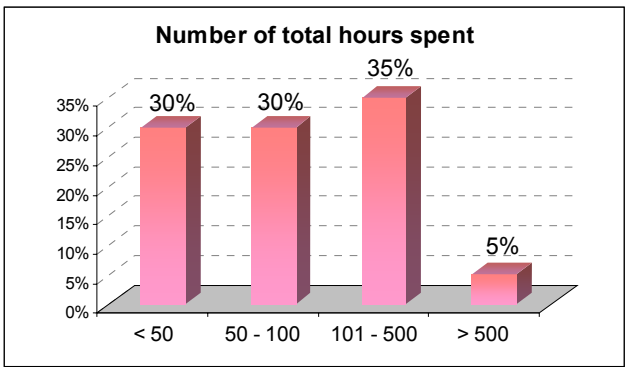


**Figure 7. Time spent by participating teams on the competition**

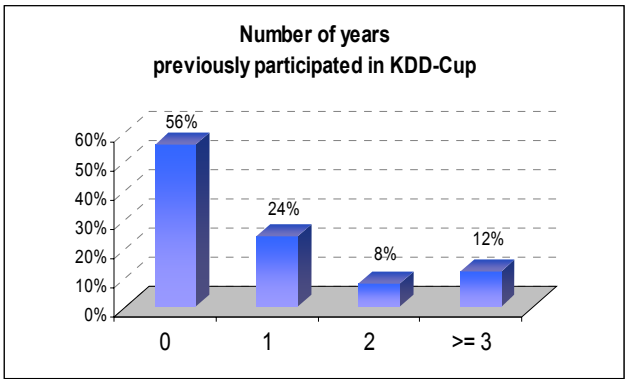


**Figure 8. New participant or return participant**

## 6. SUMMARY

The KDD-Cup 2005 task presented a real web search engine problem, categorizing search queries, to challenge the data mining community. Significant number of participants from both academy and industry and the long hours they spent indicated the great interests the community has in this area.

In this paper, we described the competition task, discussed the evaluation method, and presented the award winners. We briefly summarized the key techniques from the submitted solutions. Many of these techniques are technically significant and/or practical. Some of them achieved quite good results. To solve this task, participants developed a lot of interesting ideas. These techniques and ideas have potential to be used in practical search engines. However, given the precision and F1 scores of the submitted solutions, there is still a significant room to improve further.

To encourage the community to continue research in this area, we made all the materials of this KDD-Cup including data and

answer sets available at the KDD-Cup 2005 web site: http://www.acm.org/sigs/sigkdd/kdd2005/kddcup.html. We believe that more advanced and practical techniques for solving search problems will come from the data mining community.

## 7. ACKNOWLEDGMENTS

First we would like to thank everyone who participated in the KDD-Cup 2005 competition for their hard work and creativity. We also thank MSN Data Mining & Research team for the great help on preparing the data and evaluating the results. They are Shuzhen Nong, Jeremy Tantrum, Teresa Mah, and Abhinai Srivastava. Finally we sincerely thank KDD 2005 web master Michal Sabala for his support.

## 8. REFERENCES

[1] ACM SIGKDD 2005. http://www.acm.org/sigs/sigkdd/kdd2005.

[2] ACM SIGKDD-CUP 2005. http://www.acm.org/sigs/sigkdd/kdd2005/kddcup.html

[3] MSN Search. http://search.msn.com/

[4] C. J. van Rijsbergen, *Information Retrieval* (Second Edition). London, U.K., 1979

[5] C. D. Manning and H. Schtüze. *Foundations of Statistical Natural Language Processing*, London, U.K., 1999, 575-608.

[6] Wordnet. http://wordnet.princeton.edu/

[7] Wikipedia. http://www.wikipedia.org/

[8] C. Silverstein, M. Henzinger, H. Marais, and M. Moricz. Analysis of a very large AltaVista query log, SRC Technical Note #1998-14.

[9] B. J. Jansen and U. Pooch. A review of web searching studies and a framework for future research. *Journal of the American Society of Information Science and Technology*, 53(3):235-246, 2000.

[10] U. Lee, Z. Liu, and J. Cho. Automatic identification of user goals in web search. Technical report, UCLA Computer Science, 2004.

[11] D. E. Rose and D. Levinson. Understanding user goals in web search. In *Proc. of WWW 2004*, 2004.

[12] L. Wang, C. Wang, X. Xie, J. Forman, Y. Lu, W. Ma and Y. Li. Detecting dominant locations from search queries, In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, 2005.

[13] J. Sun, H. Zeng, H. Liu, Y. Lu, and Z. Chen. CubeSVD: A novel approach to personalized web search. In *Proceedings of the 14th international conference on World Wide Web*, 2005

[14] J. Teevan, S.T. Dumais and E. Horvitz. Personalizing search via automated analysis of interests and activities. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, 2005.

[15] Google personalized search. http://labs.google.com/personalized.

[16] My Yahoo! http://my.yahoo.com/?myhome.

## About the authors:

Ying Li is a Senior Director at MSN, Microsoft, in charge of Data Mining and Applied Research. She manages a team of applied researchers and research analysts to solve algorithmic problems for the advertising platform at MSN. Before Microsoft, she was a Technical Director at Computer Research Institute of Montreal, Canada, leading various applied research projects. She holds a Ph.D. degree in Computer Science from the University of British Columbia, Canada.

Zijian Zheng is an active data mining and machine learning researcher and practitioner. He has a computer science PhD specialized in machine learning from the University of Sydney. After that he did three years research in this area at Deakin University. In 1999, he moved to industry. Zijian has worked in Blue Martini Software, Microsoft, and Amazon.com. Where, he has analyzed different types of huge real data sets and has built commercial business intelligence software systems using data mining techniques to understand customers/users. Currently, he is in the MSN search group leading the effort of applying data mining and machine learning techniques.

Honghua (Kathy) Dai is a Data Mining Research Analyst at MSN, Microsoft. She is also a computer science Ph. D. student in DePaul University. She is actively working on applying machine learning and data mining techniques on Web / Search Data to understand Internet user behavior.