The second *Large Scale Hierarchical Text Classification* PASCAL Challenge

A. Kosmopoulos[†,◇], G. Paliouras[†], E. Gaussier[*],
I. Androutsopoulos[◇], T. Artières[‡], P. Gallinari[‡]

[*] Lab. d'Informatique de Grenoble & Grenoble University, France
[†] National Center for Scientific Research "Demokritos", Greece
[◇] Athens University of Economics and Business, Greece
[‡] Lab. d'informatique de Paris 6, France

September 1, 2011

Outline

Large scale hierarchical text classification (1)

Situation

- Problem has been addressed in the past
  - Seminal work of Yang et. al. [9] (ca. 14,000 categories) in 2003
  - Followed by extensions in 2005 ([5, 6]) to more than 100,000 categories
  - With even more categories (10^6) in the work of Beygelzimer et. all. [1] in 2009
- Comparison of different classifiers in different settings: flat vs hierarchical
- Continuous interest in the problem (under different forms) and continuous flow of new ideas and approaches - work by Xue et. al. in 2008 [8] and Bengio et. al. in 2010 [7]

Large scale hierarchical text classification (2)

Comments

- The dichotomy introduced in early works between flat and hierarchical approaches blurred in more recent works
- Different classifiers can be used differently (e.g. feature selection or document sampling/filtering can be used in flat approaches to speed up the process); experimental space is indeed large
- Recent challenges on large scale classification on large numbers of high-dimensional exmaples, but very few categories

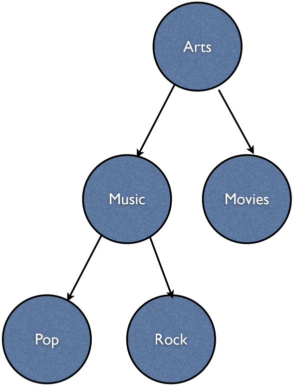
$\Rightarrow$ All these elements led us to propose a challenge on large scale hierarchical classification, the first edition of which was held in 2009/2010 [2]

What is different in the new challenge?

- The maximum number of categories increased from 12,000 to 325,000
- The maximum number of examples increased from 160,000 to 2,000,000
- We used data from Wikipedia (www.wikipedia.org), in addition to the ODP Web directory data (www.dmoz.org)
- The hierarchy of the wikipedia datasets is a graph instead of a tree
- The classification tasks are multi-label instead of single-label

Description of the Problem

- Hierarchy of categories is provided (relations of parent-child between the categories)
- Large numbers of categories (27,000 - 325,000) and examples (394,000 - 2,300,000)
- Simple hierarchical problem: documents at the leaves only
- Multi-label problem: each document can belong to more than one category



Data preparation

- Pre-processing:
  - Stemming/lemmatization
  - Stop-word removal
- Replacement of each stem with an id
- Transformation of documents into feature vectors
- Filtering of classes and documents (Wikipedia data sets)
- Splitting of documents into training and testing (not trivial for multi-label data)

Datasets

	#cat	#stems	#docs	cat/doc	max path
DMOZ	27,875	594,158	497,992	1,02	5
Wiki Small	36,504	346,299	538,148	1,86	10
Wiki Large	325,056	1,617,899	2,817,603	3.26	14

Tasks of the Challenge

Task Name	#train docs	# test docs	Hier
Task 1: Dmoz	394,756	104,263	Tree
Task 2: Wiki small	456,886	81,262	Graph
Task 3: Wiki large	2,365,436	452,167	Graph

Evaluation Measures - Example based

$$Accuracy[4] = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i \cap Z_i|}{|Y_i \cup Z_i|}$$

$$F_1[4] = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{2|Y_i \cap Z_i|}{|Y_i| + |Z_i|}$$

- D is the number of testing documents
- Z_i the predicted labels by the classifier
- Y_i the true labels of the document

Evaluation Measures - Label based

$$M_{macro}[4] = \frac{1}{|L|} \sum_{\lambda=1}^{|L|} M(tp_\lambda, fp_\lambda, tn_\lambda, fn_\lambda)$$

$$M_{micro}[4] = M(\frac{1}{|L|} \sum_{\lambda=1}^{|L|} tp_\lambda, \frac{1}{|L|} \sum_{\lambda=1}^{|L|} fp_\lambda, \frac{1}{|L|} \sum_{\lambda=1}^{|L|} tn_\lambda, \frac{1}{|L|} \sum_{\lambda=1}^{|L|} fn_\lambda))$$

$$precision = \frac{TP}{TP + FP}, \; recall = \frac{TP}{TP + FN}$$

$$F_1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

where L the labels

Evaluation Measures - Multi-label graph-induced error

$$\text{MGIE} = \frac{\sum_{d=1}^{|D|} \sum_{c}^{|min(T,P)|} \text{Min-graph-distance}(c, max(T,P))}{|\text{Classification tasks}|}$$

- D is the number of testing documents
- T the true labels of a document
- P the predicted labels of a document
- Min-graph-distance is computed in such a way that minimizes the sum of distances. In our experiments the maximum distance was set to five.
- Based on [3] but extended for multi-label data and the use of a graph instead a of tree

Significance Tests - for Macro measures

Macro sign test (S-test)[10]

$$Z = \frac{k - 0.5n}{0.5\sqrt{n}}, \text{ since } n > 12$$

- n is the number of times that a_i and b_i differ
- k is he number of times that a_i is larger than b_i
- $a_i \in [0, 1]$ is the F_1 score of system A on the ith category (i= 1, 2, ..., M)
- $b_i \in [0, 1]$ is the F_1 score of system B on the ith category (i= 1, 2, ..., M)
- M is:
  - the number of categories for label based measures
  - the number of documents for example based measures
- Significant different if P-value < 0.05

Significance Tests - for Micro measures

Micro sign test (S-test)[10]

$$Z = \frac{k - 0.5n}{0.5\sqrt{n}}, \text{ since } n > 12$$

- n is the number of times that a_i and b_i differ
- k is he number of times that a_i is larger than b_i
- $a_i \in \{0, 1\}$ is the measure of success for system A on the ith decision (i= 1, 2, ..., N)
- $b_i \in \{0, 1\}$ is the measure of success for system B on the ith decision (i= 1, 2, ..., N)
- N is the number of binary decisions
- Significant different if P-value < 0.05

Significance Tests

- The null hypothesis is that k has a binomial distribution Bin(n, p) where $p = 0.5$
  $\Rightarrow$ there is no significant difference between the two systems
- The alternative hypothesis is that he binomial distribution of k with $p > 0.5$
  $\Rightarrow$ system A is better than system B
- A larger difference doesn't always translate to significant difference
- Abnormality in significant difference between systems ranked by an evaluation measure
  For example:
  - $A > B > C$ according to evaluation measure X
  - But A appears significantly better than B but not than C

Standard Approaches

Two main approaches[8]:

Big-bang Directly categorize documents to the leaves.

Top-down Hierarchy is exploited in order to divide the problem into smaller ones.

Big-bang approaches are usually more accurate while Top-down approaches are usually faster.

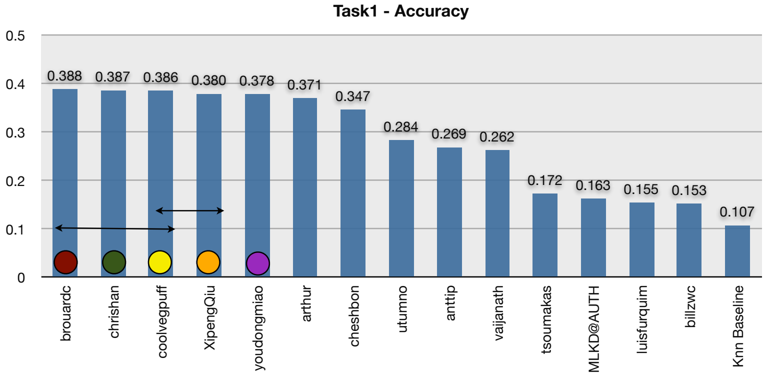
Approaches used in the Challenge

- Most participants either used a big-bang approach or exploited only a small part of the hierarchy.
- Regarding the classifiers:
  - Lazy learners were used, which are very fast at training but slower at classification. (i.e. kNN)
  - Eager learners were also used and were faster at classification. (i.e. Naïve Bayes, SVMs)

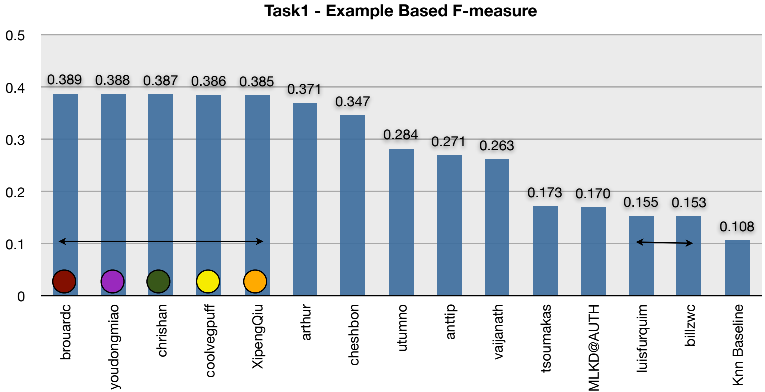
Highest Scores per Task

	Task 1	Task 2	Task 3
Accuracy	0.388	0.374	0.347
Example based F_1	0.389	0.433	0.426
Label based macro F_1	0.275	0.242	0.187
Label based micro F_1	0.389	0.390	0.348
MGIE	2.823	3.726	4.288

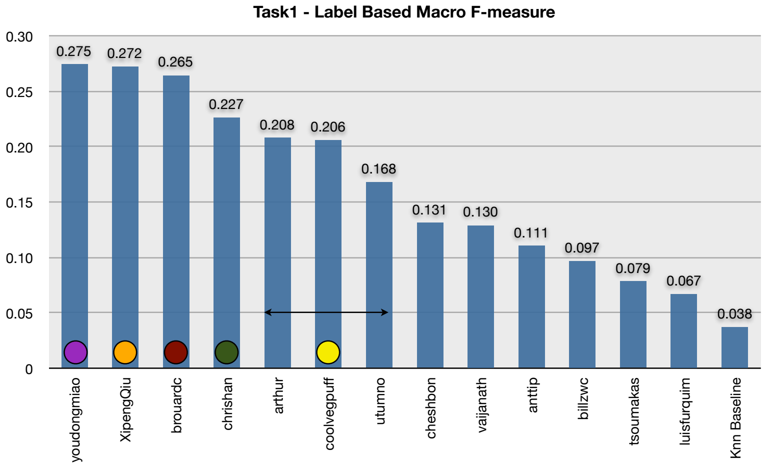
Task 1: DMOZ


Task1 - Accuracy
0.5
0.4
0.3
0.2
0.1
0
0.388
0.387
0.386
0.380
0.378
0.371
0.347
0.284
0.269
0.262
0.172
0.163
0.155
0.153
0.107
brouardc
chrishan
coolvegpuff
XipengQiu
youdongmiao
arthur
cheshbon
utumno
anttip
vaijanath
tsoumakas
MLKD@AUTH
luisfurquim
billzwc
Knn Baseline

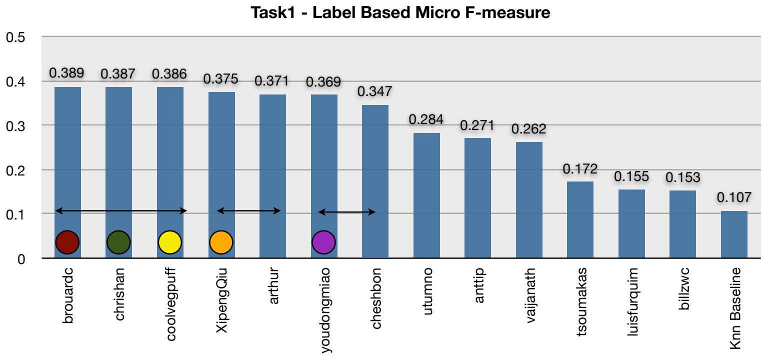
Task 1: DMOZ


Task1 - Example Based F-measure
0.5
0.4
0.3
0.2
0.1
0
0.389
0.388
0.387
0.386
0.385
0.371
0.347
0.284
0.271
0.263
0.173
0.170
0.155
0.153
0.108
brouardc
youdongmiao
chrishan
coolvegpuff
XipengQiu
arthur
cheshbon
utumno
anttip
vaijanath
tsoumakas
MLKD@AUTH
luisfurquim
billzwc
Knn Baseline

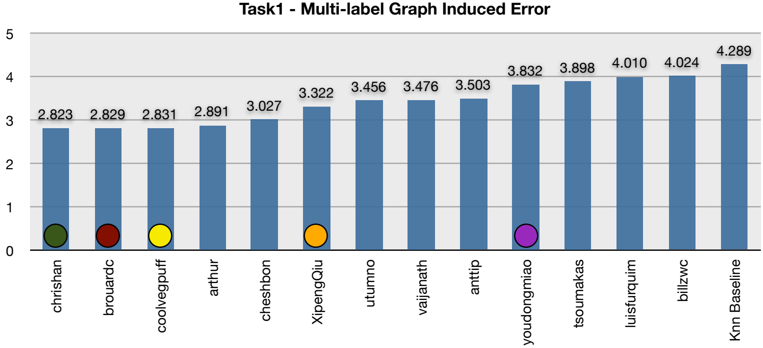
Task 1: DMOZ


Task1 - Label Based Macro F-measure
0.30
0.25
0.20
0.15
0.10
0.05
0
0.275
0.272
0.265
0.227
0.208
0.206
0.168
0.131
0.130
0.111
0.097
0.079
0.067
0.038
youdongmiao
XipengQiu
brouardc
chrishan
arthur
coolvegpuff
utumno
cheshbon
vaijanath
anttip
billzwc
tsoumakas
luisfurquim
Knn Baseline

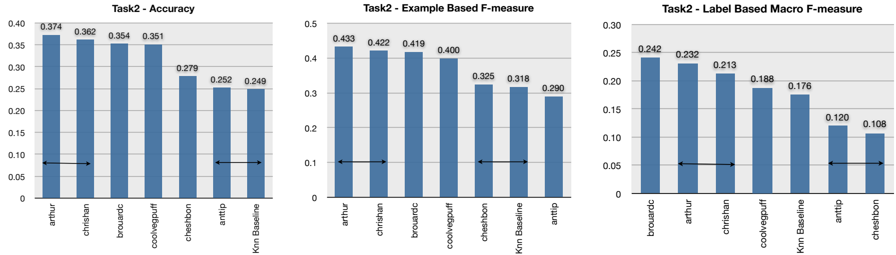
Task 1: DMOZ


Task1 - Label Based Micro F-measure
0.5
0.4
0.3
0.2
0.1
0
0.389
0.387
0.386
0.375
0.371
0.369
0.347
0.284
0.271
0.262
0.172
0.155
0.153
0.107
brouardc
chrishan
coolvegpuff
XipengQiu
arthur
youdongmiao
cheshbon
utumno
anttip
vaijanath
tsoumakas
luisfurquim
billzwc
Knn Baseline

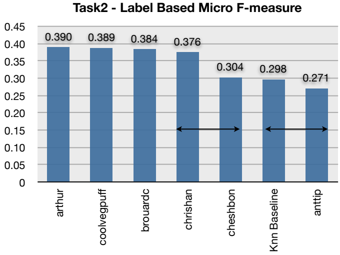
Task 1: DMOZ

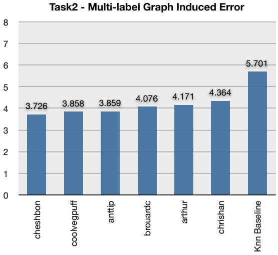

Task1 - Multi-label Graph Induced Error
5
4
3
2
1
0
2.823
2.829
2.831
2.891
3.027
3.322
3.456
3.476
3.503
3.832
3.898
4.010
4.024
4.289
chrishan
brouardc
coolvegpuff
arthur
cheshbon
XipengQiu
utumno
vaijanath
anttip
youdongmiao
tsoumakas
luisfurquim
billzwc
Knn Baseline

Task 2: Wikipedia small

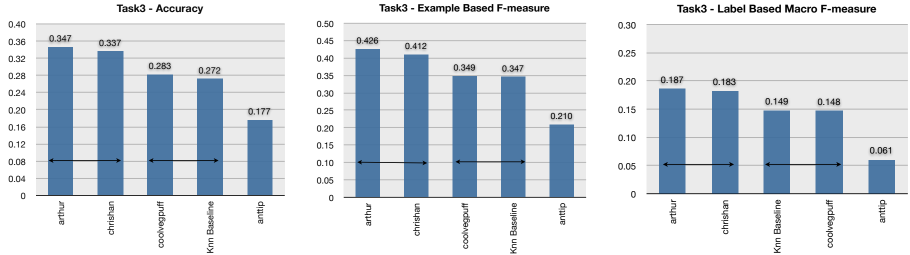


Task 2: Wikipedia small

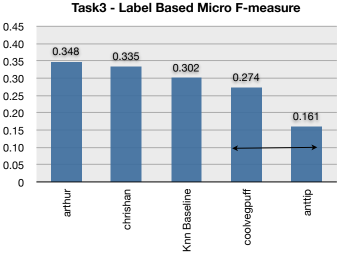


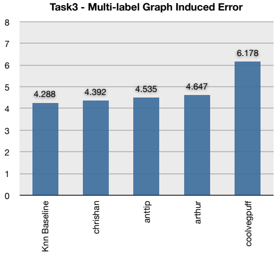


Task 3: Wikipedia large



Task 3: Wikipedia large





Conclusion and Perspectives

- All the approaches we are aware of on large scale classification tried by participants; we thus believe the results obtained represent the state-of-the-art on this collection
- No complete hierarchical approaches, a la pachinko; rather approaches with a limited use of the hierarchy
- Usual evaluation measures are not ideal in hierarchical classification
- Are the significant tests appropriate for this problem; what else should be used?
- Useful benchmark for future use; oracle is available at the challenge site
- LSHTC-3 - What should be different? Original text? Other data except text?

Y. Sorkin A. Beygelzimer, J. Langford and A. Stehl.
Conditional probability tree estimation analysis and algorithms.

In *UAI*, 2009.

G. Paliouras A. Kosmopoulos, E. Gaussier and Aseervatham.
The ECIR 2010 large scale hierarchical classification workshop.
In *SIGIR Forum*, volume 44, pages 23–32, 2010.

Ofer Dekel, Joseph Keshet, and Yoram Singer.
Large margin hierarchical classification.
In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, page 27, New York, NY, USA, 2004. ACM.

I. Katakis G. Tsoumakas and I. Vlahavas.
Random k-labelsets for multi-label classification.

In *IEEE Transactions on Knowledge Discovery and Data Engineering*, 2010.

Tie-Yan Liu, Yiming Yang, Hao Wan, Hua-Jun Zeng, Zheng Chen, and Wei-Ying Ma.
Support vector machines classification with a very large-scale taxonomy.
SIGKDD Explorations, 7(1):36–43, 2005.

Tie-Yan Liu, Yiming Yang, Hao Wan, Qian Zhou, Bin Gao, Hua-Jun Zeng, Zheng Chen, and Wei-Ying Ma.
An experimental study on large-scale web categorization.
In Allan Ellis and Tatsuya Hagino, editors, *WWW (Special interest tracks and posters)*, pages 1106–1107. ACM, 2005.

J. Weston S. Bengio and D. Grangier.
Label embedding trees for large multi-class tasks.
In *NIPS*, 2010.

Gui-Rong Xue, Dikan Xing, Qiang Yang, and Yong Yu.
Deep classification in large-scale text hierarchies.
In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 619–626, New York, NY, USA, 2008. ACM.

B. Kisiel Y. Yang, J. Zhang.
A scalability analysis of classifiers in text.
In *ACM SIGIR Conference*. ACM, 2003.

Yiming Yang and Xin Liu.
A re-examination of text categorization methods.
pages 42–49. ACM Press, 1999.