Иерархическая классификация текстов. Конкурс The Large Scale Hierarchical Text Classification

Остапец Андрей

19 ноября 2014 г.

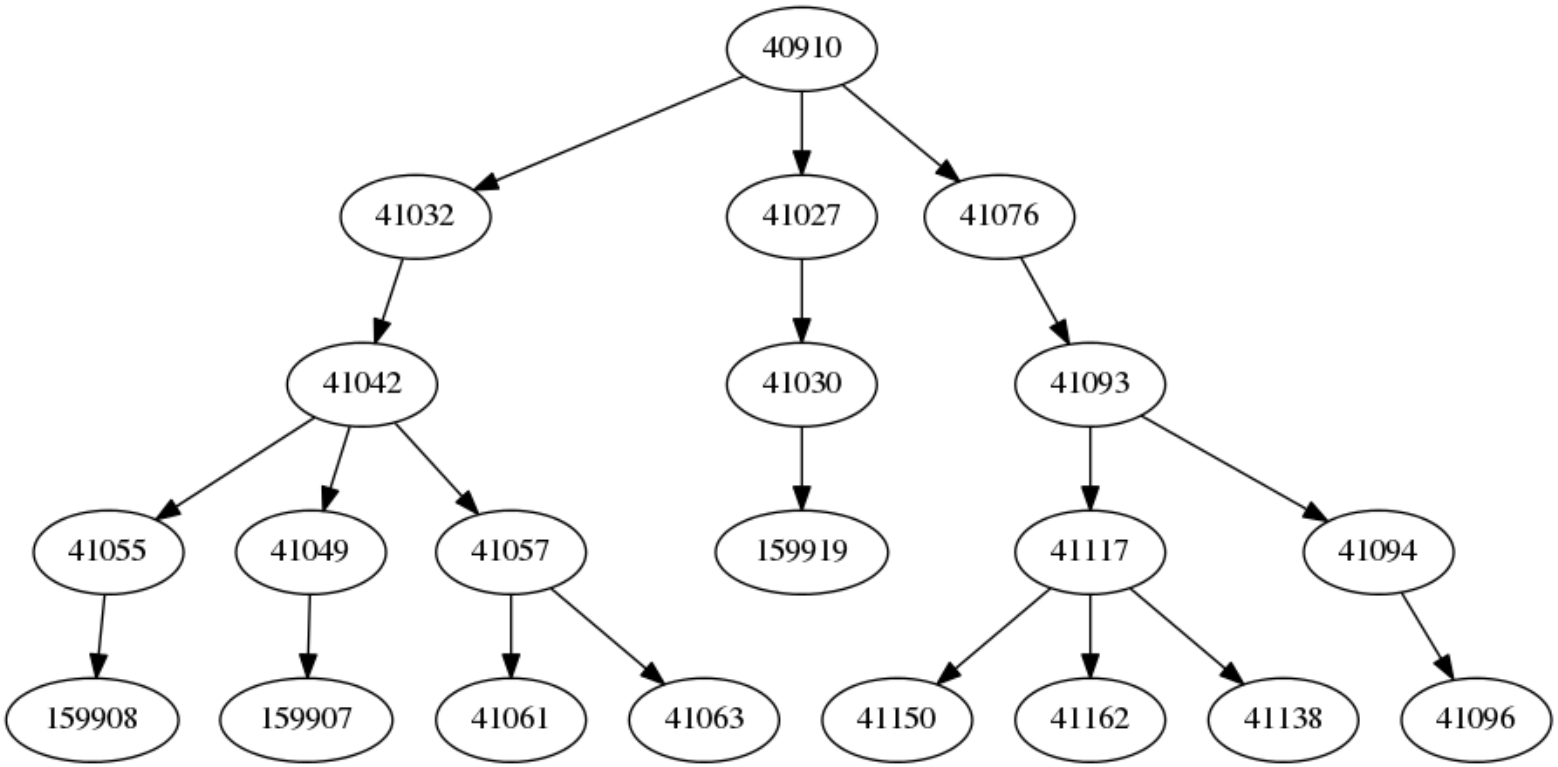
Сегодня поговорим о...

- Иерархической классификация текстов
- Конкурсе LSHTC1
- Лучших решениях в этом конкурсе

Постановка задачи

- Имеется множество документов D и множество классов C, организованных в иерархию, каждому документу из D приписан один класс из C.
- Требуется на основе этих данных построить процедуру автоматической классификации текстов.
- Структуру классов представляет собой дерево, причем классификация проходит только по листьям этого дерева.

Иерархия категорий



История

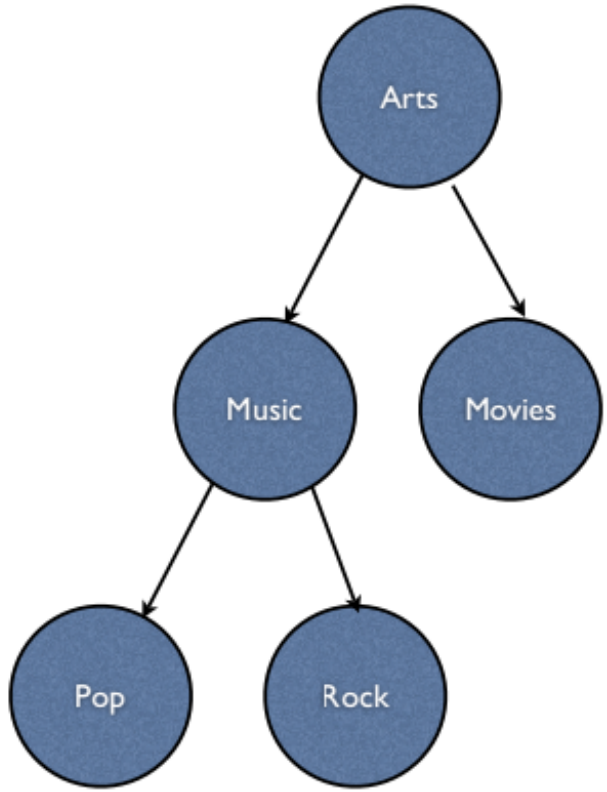
- Впервые задача была представлена в 2003 году (14 000 категорий) [1].
- В 2005 году задача классификации на более чем 100 000 категорий [2], [3].
- Два основных подхода: если алгоритм учитывает иерархию классов, он называется иерархическим (hierarchical), если не учитывает – плоским (flat).

Настоящее время

- Строгое разделение между двумя этими подходами, представленное в первых работах, исчезает в последующих работах.
- Последние соревнования - большое число документов, признаковое пространство высокой размерности, но растет число документов на одну категорию.

Описание конкурса LSHTC1

- Проводился с июля по декабрь 2009 года.
- Данные проекта являются частью Открытого Каталога (ODP): http://www.dmoz.org/
- Задание состоит из четырех блоков данных, данные в блоках частично повторяются.
- Простая иерархия: документы могут принадлежать только «листовому» классу в иерархическом дереве.
- Каждый документ может принадлежать только одному классу.



Подготовка данных

- Предобработка (стемминг/лемматизация)
- Удаление стоп-слов
- Два типа векторов:

Content data

The Movies Net Top 20 brings you the pick of the most popular and highest-rating sites for movies on the Net today. It gives you access to the world's best movies sites, all from a single page.

Movies.com features the latest movie news, reviews, trailers and a wide variety of general movie information.

Description data

Movies Top 20 - Lists links to several film sites, with brief descriptions of each.

Наборы данных

- Большой набор данных (12294 категорий)
  Использовался для оценки результатов
- Маленький набор данных (1139 категорий)
  Предложен участникам для настройки алгоритмов
- Каждый набор данных разделен на:
  - обучение (93805/4463 документов)
  - валидацию (34905/1860 документов)
  - тест (34880/1858 документов)

4 задания конкурса

Task Name	Content		Description	
	Train	Test	Train	Test
Task 1: Basic	✓	✓	-	-
Task 2: Cheap	-	✓	✓	-
Task 3: Expensive	✓	✓	✓	-
Task 4: Full	✓	✓	✓	✓

Число признаков в каждом задании

Task Name	Train	Validation	Test
Task 1: Basic	347255	191224	194024
Task 2: Cheap	71322	39070	194024
Task 3: Expensive	368113	201487	194024
Task 4: Full	368113	201487	204288

Оценка результатов

$$\text{Accuracy} = \frac{m}{D}$$

$$\text{Tree-induced error} = \frac{\sum_{d=1}^{M} \text{Path-length}(c_d, t_d)}{M}$$

- D - количество документов в тесте.
- m - количество правильно классифицированных документов.
- c_d - категория определенная для документа d классификатором
- t_d - истинная категория документа

Оценка результатов

$$\text{Macro precision} = \frac{\sum_{i=1}^{M} \text{precision}_i}{M}, \ \text{precision} = \frac{|u \cap v|}{|u|}$$

$$\text{Macro recall} = \frac{\sum_{i=1}^{M} \text{recall}_i}{M}, \ \text{recall} = \frac{|u \cap v|}{|v|}$$

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

- v - множество документов, действительно принадлежащих категории
- u - множество документов, приписанных категории алгоритмом
- M - количество категорий

Micro меры

$$\text{Micro precision} = \frac{\sum_{i=1}^{M} |u_i \cap v_i|}{\sum_{i=1}^{M} |u_i|},$$

$$\text{Micro recall} = \frac{\sum_{i=1}^{M} |u_i \cap v_i|}{\sum_{i=1}^{M} |v_i|},$$

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

- v_i - множество документов, действительно принадлежащих категории c_i
- u_i - множество документов, приписанных категории c_i алгоритмом
- M - количество категорий

Стандартные подходы

Два основных подхода:

- **Big-bang** Непосредственно классифицируем документы по узлам.
- **Top-down** Последовательно спускаем по дереву иерархии, каждый раз делим задачу на несколько новых меньшего размера

Первый подход обычно более точный, второй подход обычно быстрее.

Подходы участников

- Большинство участников либо вообще не использовали иерархию, либо использовали только маленькую часть дерева.
- Отбор признаков производился, но не всегда помогал
- Участники, которые учитывали иерархию, использовали информацию либо двух верхних слоев, либо предпоследнего.

Результаты

Task 1: Basic Evaluation Results

Results Ordered by Accuracy:

Name	Accuracy	Macro F-measure	Macro Precision	Macro Recall	Tree Induced Error
alpaca	0.467632	0.341244	0.323491	0.361059	3.07858
jhuang	0.463217	0.35494	0.332977	0.380005	3.28079
arthur_general	0.443291	0.31965	0.303413	0.337723	3.29289
XipengQiu	0.443062	0.336739	0.314129	0.362856	3.36307
Turing	0.43168	0.323213	0.304554	0.344306	3.39343
Dyakonov	0.42695	0.328417	0.304365	0.356597	3.43323
logicators	0.415224	0.328591	0.296748	0.36809	3.57964
ysyky	0.408372	0.296651	0.272088	0.32609	3.4947
zwner	0.405218	0.28168	0.268347	0.296407	3.55427
wdd	0.403326	0.294133	0.280493	0.309167	3.57761
NakaCristo	0.402322	0.286546	0.268028	0.307813	3.65049
shawndr	0.401806	0.315507	0.292382	0.342604	3.71531
alfonsoeromero	0.361325	0.285235	0.259169	0.317131	4.08865
bbutc	0.36035	0.22574	0.20559	0.250271	4.40313
illes.solt	0.314708	0.199466	0.19709	0.2019	4.43045
leokury	0.31078	0.170418	0.171364	0.169483	4.24974
semelak1	0.301175	0.247711	0.231118	0.266872	4.69521
brouardc	0.291714	0.185995	0.162741	0.217004	4.48804
controlledchaos	0.151978	0.135923	0.111545	0.173937	6.38271

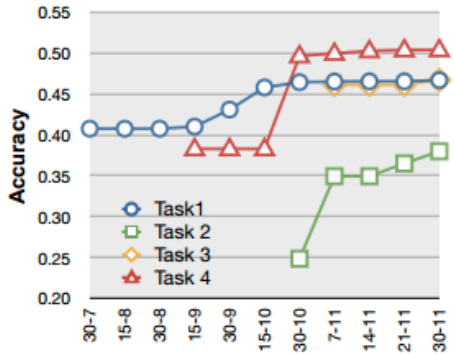
Подходы участников

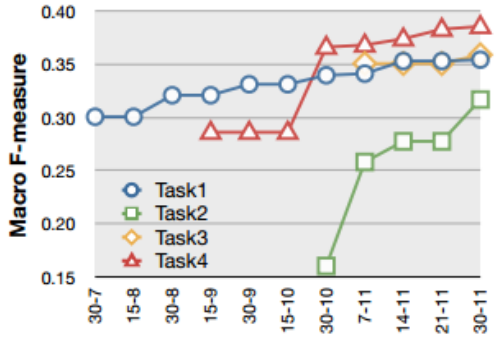
- alpaca - комбинация классификаторов, SVM с полиномиальным ядром второй степени.
- jhuang - плоский классификатор, два онлайн алгоритма (OOZ и PA).
- arthur general - иерархия на двух уровнях, многоклассовый SVM.
- XipengQiu - центроид для каждого класса.
- Turing - KNN для поиска подмножества вероятных кандидатов и затем наивный байесовский классификатор.
- logicators - подмножество иерархии, иерархический SVM.
- NakaCristo - плоский классификатор, взвешенный метод ближайшего соседа
- Brouard - классификация с помощью резонанса ассоциативной сети

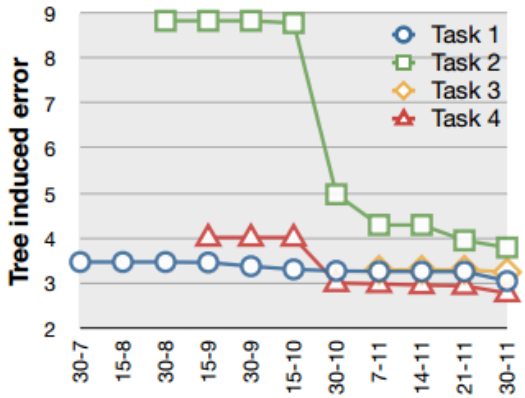
Лучшие результаты по каждому заданию

	Task 1	Task 2	Task 3	Task 4
Accuracy	0.467632	0.380619	0.467861	0.504759
Macro F-measure	0.35494	0.317133	0.359557	0.386195
Tree Induced Error	3.07858	3.80803	3.2621	2.82101

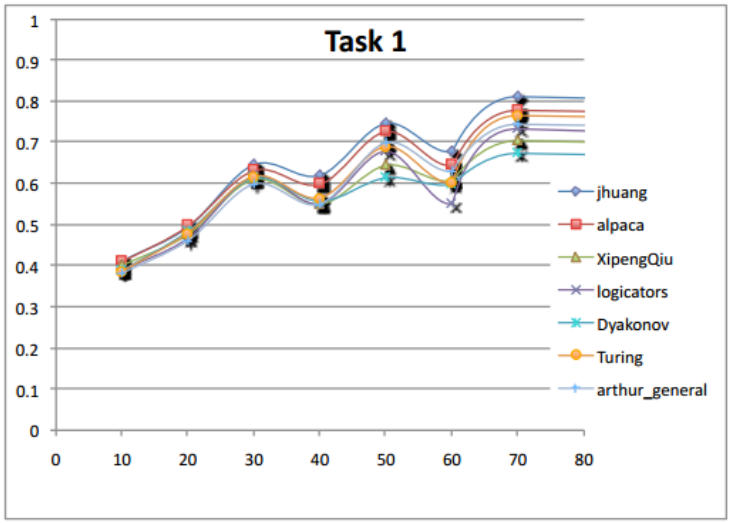
Изменение лучших результатов

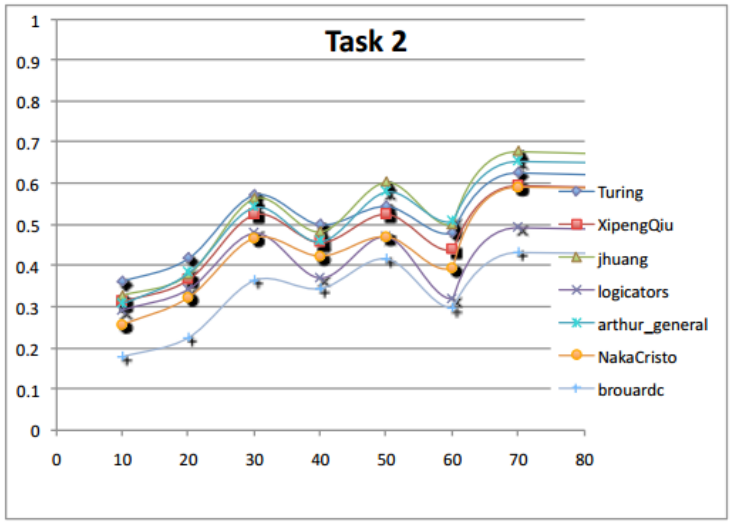


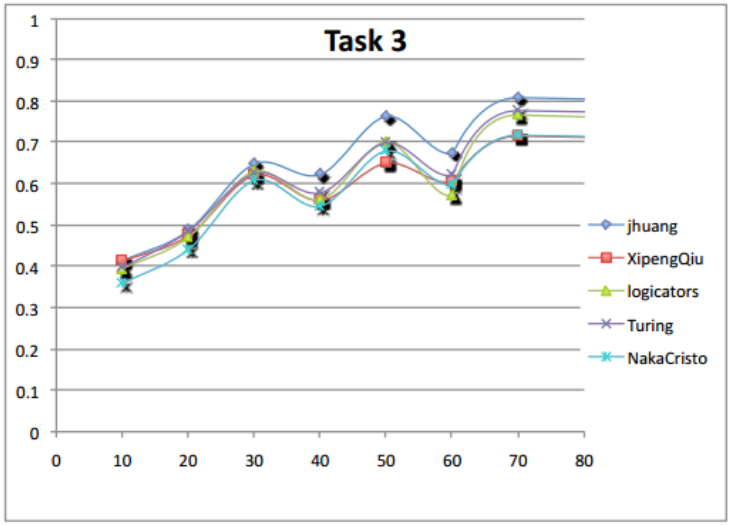


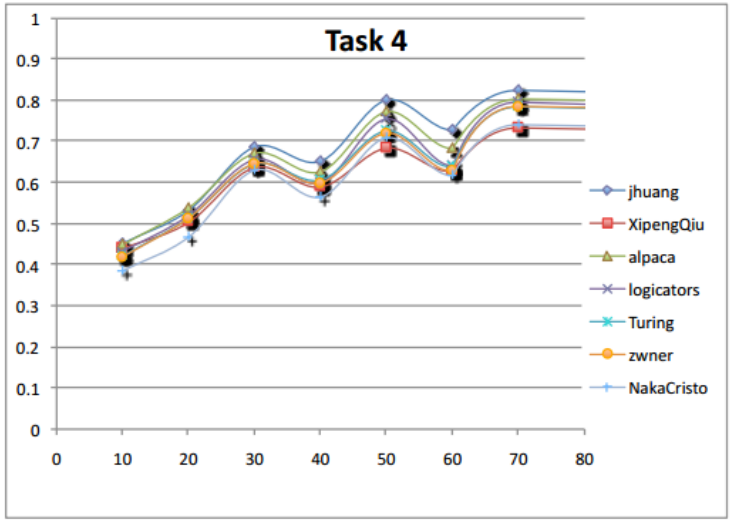


F1-мера от размера категории









Время работы

Categories	XipengQiu	logicators	Turing	NakaCristo
12294	1m 12.5s 94m 8.2s	**67m 1.4s** **296m 45s**	**0m 13s** **1258m 2s**	18.8s 42m 5s
10000	0m 58s 60m 3s	41m 20.7s 153m 20s	0m 8s 779m 35s	4m 11.5s 27m 6s
1000	0m 7.6s 0m 42s	0m 35s 1m 28s	0m 0.7s 9m 36s	0m 24.2s 0m 30.2s
100	0m 4.7s 0m 1.2s	0m 0.2s 0m 3.7s	0m 13s 0m 22.6s	0m 2.5s 0m 1.9s

Первое время - обучение

Второе время - классификация

Память

Categories	XipengQiu	logicators	Turing	NakaCristo
12294	2920 Mb 1382 Mb	**5700 Mb** **3900 Mb**	**200 Mb** **6000 Mb**	921 Mb 996 Mb
10000	2400 Mb 1050 Mb	4600 Mb 2950 Mb	76 Mb 5800 Mb	828 Mb 762 Mb
1000	170 Mb 149 Mb	444 Mb 434 Mb	$<$ 50 Mb 1320 Mb	110Mb 79 Mb

Первое время - обучение
Второе время - классификация

Литература


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


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.