Experimenting SnakeT: A Hierarchical Clustering Engine for Web-Page Snippets*

Paolo Ferragina and Antonio Gullì

Dipartimento di Informatica, Università di Pisa
{ferragina,gulli}@di.unipi.it

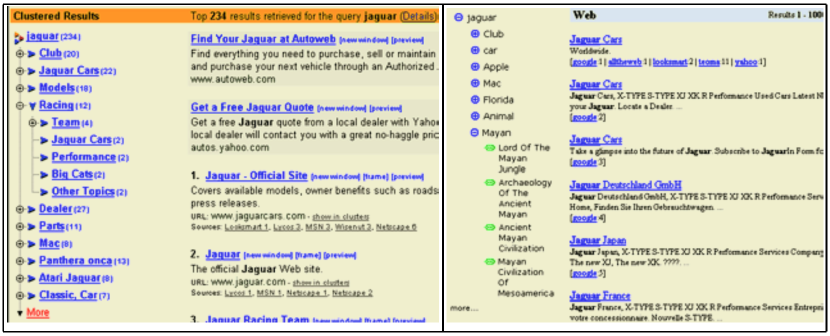
Current search engines return a ranked list of web pages represented by page excerpts called the *web snippets*. The ranking is computed according to some relevance criterium that takes into account textual and hyperlink information about the web pages (see e.g. [1]). This approach is very well-known and a lot of research is pushing towards the design of better and faster ranking criteria. However, it is nowadays equally known that a flat list of results limits the retrieval of *precise* answers because of many factors. First, the relevance of the query results is a subjective and time-varying concept that strictly depends on the context in which the user is formulating the query. Second, the ever growing web is enlarging the number and heterogeneity of candidate query answers. Third, the web users have limited patience so that they usually just look at the top ten results. The net outcome of this scenario is that the retrieval of the correct answer by a standard user is getting more and more difficult, if not impossible.

It is therefore not surprising that new IR tools are being designed to *boost*, or *complement*, the efficacy of search-engine ranking algorithms. These tools offer new ways of organizing and presenting the query results that are more intuitive and simple to be browsed, so that the users may match their needs faster. Among the various proposals, one became recently popular thanks to the engine Vivisimo (see Figure 1) that got in the last three years the "Best Metasearch Engine Award" by SearchEngineWatch.com.

The goal of this demonstration proposal is to describe the functioning of a Web Hierarchical Clustering engine developed at the University of Pisa, called SnakeT. The algorithmic ideas underlying this IR-tool has been briefly described in a companion paper published in this proceedings. SnakeT offers some distinguishing features with respect to known solutions that make it much closer to Vivisimo's results: (1) it offers a large coverage of the web by drawing the snippets from 16 search engines (e.g. Google, MSN, Overture, Teoma, and Yahoo), the Amazon collection of books (a9.com) and the Google News, in a flexible and efficient way via I/O Async; (2) it builds the clusters and their labels on-the-fly in response to a user query (without using any predefined taxonomy); (3) it selects on-the-fly the best labels by exploiting a variant of the TF-IDF measure computed onto the whole web directory DMOZ; (4) it organizes the clusters and their labels in a hierarchy, by minimizing an objective function which takes into account various features that abstract some quality and quantitative requirements.

* Partially supported by the Italian MIUR projects ALINWEB and ECD, and by the Italian Registry of ccTLD.it.

J.-F. Boulicaut et al. (Eds.): PKDD 2004, LNAI 3202, pp. 543–545, 2004.



Clustered Results — Top **234** results retrieved for the query **jaguar** (Details)

- jaguar (234)
  - Club (20)
  - Jaguar Cars (22)
  - Models (18)
  - Racing (12)
    - Team (4)
    - Jaguar Cars (2)
    - Performance (2)
    - Big Cats (2)
    - Other Topics (2)
  - Dealer (27)
  - Parts (11)
  - Mac (8)
  - Panthera onca (13)
  - Atari Jaguar (8)
  - Classic, Car (7)
  - More

Find Your Jaguar at Autoweb [new window] [preview]
Find everything you need to purchase, sell or maintain and purchase your next vehicle through an Authorized .
www.autoweb.com

Get a Free Jaguar Quote [new window] [preview]
Get a free **Jaguar** quote from a local dealer with Yahoo local dealer will contact you with a great no-haggle pric
autos.yahoo.com

1. Jaguar - Official Site [new window] [frame] [preview]
Covers available models, owner benefits such as roads press releases.
URL: www.jaguarcars.com - show in clusters
Sources: Looksmart 1, Lycos 3, MSN 3, Wisenut 2, Netscape 6

2. Jaguar [new window] [frame] [preview]
The official **Jaguar** Web site.
URL: www.jaguar.com - show in clusters
Sources: Lycos 1, MSN 1, Netscape 1, Netscape 2

3. Jaguar Racing Team [new window] [frame] [preview]

- jaguar
  - Club
  - car
  - Apple
  - Mac
  - Florida
  - Animal
  - Mayan
    - Lord Of The Mayan Jungle
    - Archaeology Of The Ancient Mayan
    - Ancient Mayan Civilization
    - Mayan Civilization Of Mesoamerica
- more...

Web Results 1 - 100

Jaguar Cars
Worldwide.
[google 1 | alltheweb 1 | looksmart 2 | teoma 11 | yahoo 1]

Jaguar Cars
Jaguar Cars, X-TYPE S-TYPE XJ XK R Performance Used Cars Latest N
your **Jaguar**. Locate a Dealer. ...
[google 2]

Jaguar Cars
Take a glimpse into the future of **Jaguar**. Subscribe to **Jaguar**In Form fo
[google 3]

Jaguar Deutschland GmbH
Jaguar Deutschland GmbH, X-TYPE S-TYPE XJ XK R Performance Serv
Home, Finden Sie Ihren Gebrauchtwagen. ...
[google 4]

Jaguar Japan
Jaguar Japan, X-TYPE S-TYPE XJ XK R Performance Services Company
The new XJ, The new XK. ????. ...
[google 5]

Jaguar France
Jaguar France, X-TYPE S-TYPE XJ XK R Performance Services Entrepri
votre concessionnaire. Nouvelle S-TYPE. ...

Fig. 1. Two Hierarchical Clustering Engines– Vivisimo and SnakeT– tested on the ambiguous query "*Jaguar*".

Another specialty of our software is that the labels are *non* contiguous sequences of terms drawn from the snippets. This offers much power in capturing meaningful labels from the poor content of the snippets (see the future works of [7, 6]). The web interface of SnakeT is available at `http://roquefort.di.unipi.it/`.

Here we briefly comment on the latest results [5, 8, 4] for comparison. [5] extracts meaningful sentences from a snippet by using a pre-computed language model, and builds the hierarchy via a recursive algorithm. The authors admit that their hierarchies are often non compact, have large depth and contain some non content-bearing words which tend to repeat. Our work aims at overcoming these limitations by ranking non-contiguous sentences and using a novel covering algorithm for compacting the hierarchies. [8] extracts variable length (contiguous) sentences by combining five different measures through regression. Their clustering is flat, and thus the authors highlight the need of (1) a hierarchical clustering for more efficient browsing, and (2) external taxonomies for improving labels precision. This is actually what we do in our hierarchical engine by developing a *ranker* based on the whole DMOZ. [4] proposes a greedy algorithm to build the hierarchy based on a minimization of an objective function similar to ours. However, their labels are contiguous sentences and usually consist of single words. The best results to date are "Microsoft and IBM products" [4, 8] not publicly accessible, as the authors communicated to us!

In the demo we will discuss the system and comment on some user studies and experiments aimed at evaluating its efficacy. Following the approach of [5], we have selected few titles from TREC Topics and used them to create a testbed enriched with other well-know ambiguous queries. See Figure 2. Following [2, 3] we then prepared two user studies. The first study was aimed at understanding whether a Web clustering engine is an useful complement to the flat, ranked list of search-engine results. We asked to 45 people, of intermediate web ability, to use Vivisimo during their day by day search activities. After a test period of 20 days, 85% of them reported that using the tool "[..] get a good sense of range alternatives with their meaningful labels", and 72% of them reported that one of the most useful feature is "[..] the ability to produce on-the-fly clusters

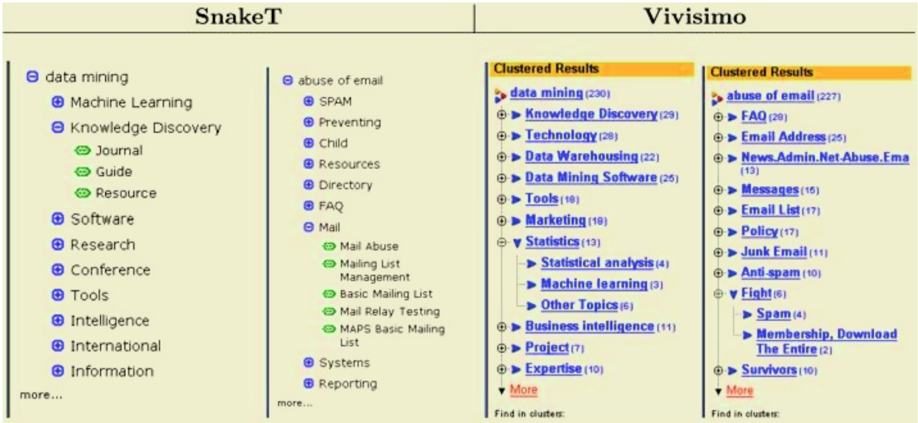


Fig. 2. Two examples of queries: *Data Mining* and *Abuse of Email*, executed on SnakeT and Vivisimo.

in response to a query, with labels extracted from the text". The second study was aimed at drawing a preliminary evaluation of our software. We selected 20 students of the University of Pisa, each of whom executed 20 different queries drawn from the testbed above. The participants evaluated the answers provided by our hierarchical clustering engine with respect to Vivisimo. 75% of them were satisfied of the quality of our hierarchy and its labels. This evaluation exercise is going on by increasing the number of users, allowing free queries, and extending its use to other domains, like blogs, news and books.

References

1. S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.
2. H. Chen and S. T. Dumais. Bringing order to the web: automatically categorizing search results. In *SIGCHI-00*, pages 145–152, 2000.
3. M. A. Hearst and J. O. Pedersen. Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In *SIGIR-96*, pages 76–84, Zürich, CH, 1996.
4. K. Kummamuru, R. Lotlikar, S. Roy, K. Singal, and R. Krishnapuram. A hierarchical monothetic document clustering algorithm for summarization and browsing search results. In *WWW*, 2004.
5. D. J. Lawrie and W. B. Croft. Generating hiearchical summaries for web searches. In *ACM SIGIR*, pages 457–458, 2003.
6. Y. S. Maarek, R. Fagin, I. Z. Ben-Shaul, and D. Pelleg. Ephemeral document clustering for web applications. Technical Report RJ 10186, IBM Research, 2000.
7. O. Zamir and O. Etzioni. Grouper: a dynamic clustering interface to Web search results. *Computer Networks*, 31:1361–1374, 1999.
8. H. Zeng, Q. He, Z. Chen, and W. Ma. Learning to cluster web search results. In *ACM SIGIR*, 2004.