

Lycos Retriever: An Information Fusion Engine

Brian Ulicny

Versatile Information Systems, Inc.
5 Mountainview Drive
Framingham, MA 01701 USA
bulicny@vistology.com

Abstract

This paper describes the Lycos Retriever system, a deployed system for automatically generating coherent topical summaries of popular web query topics.

1 Introduction

Lycos Retriever¹ is something new on the Web: a patent-pending *information fusion engine*. That is, unlike a search engine, rather than returning ranked documents links in response to a query, Lycos Retriever categorizes and disambiguates topics, collects documents on the Web relevant to the disambiguated sense of that topic, extracts paragraphs and images from these documents and arranges these into a coherent summary report or background briefing on the topic at something like the level of the first draft of a Wikipedia² article. These topical pages are then arranged into a browsable hierarchy that allows users to find related topics by browsing as well as searching.

2 Motivations

The presentation of search results as ranked lists of document links has become so ingrained that it is hard now to imagine alternatives to it. Other interfaces, such as graphical maps or visualizations, have not been widely adopted. Question-answering interfaces on the Web have not had a high adoption

rate, either: it is hard to get users to venture beyond the 2.5 word queries they are accustomed to, and if question-answering results are not reliably better than keyword search, users quickly return to keyword queries. Many user queries specify nothing more than a topic anyway.

But why treat common queries exactly like unique queries? For common queries we know that incentives for ranking highly have led to techniques for artificially inflating a site's ranking at the expense of useful information. So the user has many useless results to sift through. Furthermore, users are responsive to filtered information, as the upsurge in popularity of Wikipedia and Answers.com demonstrate.

Retriever responds to these motivations by automatically generating a narrative summary that answers, "What do I need to know about this topic?" for the most popular topics on the Web.³

3 Lycos Retriever pages

Figure 1 shows a sample Retriever page for the topic "Mario Lemieux".⁴ The topic is indicated at the upper left. Below it is a category assigned to the topic, in this case *Sports > Hockey > Ice Hockey > National Hockey League > Lemieux, Mario*. The main body of the page is a set of paragraphs beginning with a biographical paragraph complete with Lemieux's birth date, height, weight and position extracted from Nationmaster.com, followed by paragraphs outlining his career from

¹ <http://www.lycos.com/retriever.html>. Work on Retriever was done while author was employed at Lycos.

² <http://www.wikipedia.org>

³ See (Liu, 2003) for a similarly motivated system.

⁴ For other categories, see e.g. King Kong (1933):

<http://www.lycos.com/info/king-kong-1933.html>,

Zoloff: <http://www.lycos.com/info/zoloff.html>,

Public-Key Cryptography: <http://www.lycos.com/info/public-key-cryptography.html>,

Lyme Disease: <http://www.lycos.com/info/lyme-disease.html>,

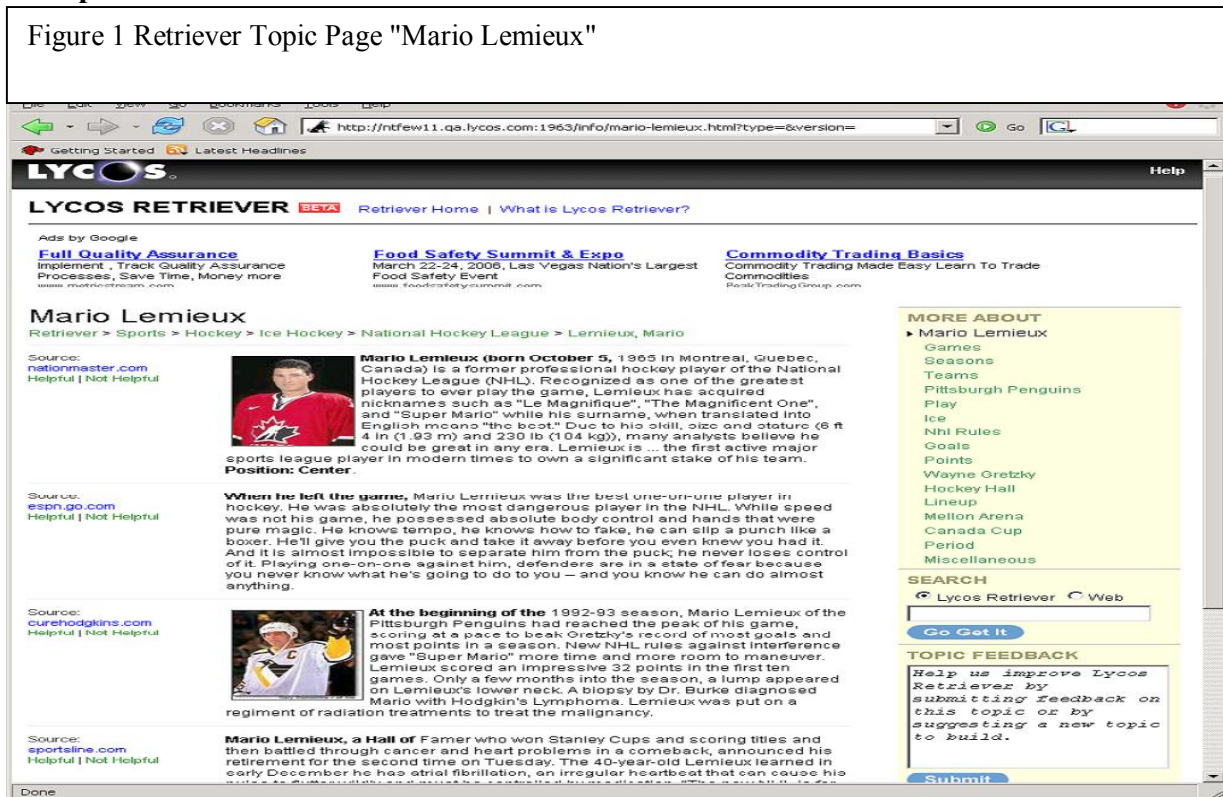
Reggaeton: <http://www.lycos.com/info/reggaeton.html>

other sources. The source for each extract is indicated in shortened form in the left margin of the page; mousing over the shortened URL reveals the full title and URL. Associated images are thumbnailed alongside the extracted paragraphs.

Running down the right side of the page under *More About* is a set of subtopics. Each subtopic is a link to a page (or pages) with paragraphs about the topic (Lemieux) with respect to such subtopics as *Games*, *Seasons*, *Pittsburgh Penguins*, *Wayne Gretzky*, and others, including the unpromising subtopic *ice*.

4 Topic Selection

Figure 1 Retriever Topic Page "Mario Lemieux"



An initial run of about 60K topics was initiated in December, 2005; this run yielded approximately 30K Retriever topic pages, each of which can have multiple display pages. Retriever topics that had fewer than three paragraphs or which were categorized as pornographic were automatically deleted. The biggest source of topic candidates was Lycos's own query logs. A diverse set of topics was chosen in order to see which types of topics generated the best Retriever pages.

5 Topic Categorization & Disambiguation

After a topic was input to the system, the Retriever system assigned it a category using a naïve Bayes classifier built on a spidered DMOZ⁵ hierarchy. Various heuristics were implemented to make the returned set of categories uniform in length and depth, up-to-date, and readable.

Once the categorizer assigned a set of categories to a topic, a disambiguator module determined whether the assigned categories could be assigned to a single thing using a set of disambiguating features learned from the DMOZ data itself. For example, for the topic 'Saturn', the assigned categories included 'Science/Astronomy', 'Recrea-

tion/Autos' and 'Computers/Video Games' (Sega Saturn). The disambiguator detected the presence of feature pairs in these that indicated more than one topic. Therefore, it clustered the assigned categories into groups for the car-, astronomy- and video-game-senses of the topic and assigned each group a discriminative term which was used to disambiguate the topic: *Saturn (Auto)*, *Saturn (Solar System)*, *Saturn (Video Game)*. Retriever returned pages only for topics that were believed to be disambiguated according to DMOZ. If no categories

⁵ <http://www.dmoz.com>

were identified via DMOZ, a default *Other* category was assigned unless the system guessed that the topic was a personal name, based on its components.

The live system assigns non-default categories with 86.5% precision; a revised algorithm achieved 93.0% precision, both based on an evaluation of 982 topics. However, our precision on identifying unambiguous topics with DMOZ was only 83%. Still, this compares well with the 75% precision achieved on by the best-performing system on a similar task in the 2005 KDD Cup (Shen 2005).

6 Document Retrieval

After a topic was categorized and disambiguated, the disambiguated topic was used to identify up to 1000 documents from Lycos' search provider. For ambiguous topics various terms were added as optional 'boost' terms, while terms from other senses of the ambiguous topic categories were prohibited. Other query optimization techniques were used to get the most focused document set, with non-English and obscene pages filtered out

7 Passage Extraction

Each URL for the topic was then fetched. An HTML parser converted the document into a sequence of contiguous text blocks. At this point, contiguous text passages were identified as being potentially interesting if they contained an expression of the topic in the first sentence.

When a passage was identified as being potentially interesting, it was then fully parsed to see if an expression denoting the topic was the Discourse Topic of the passage. Discourse Topic is an under-theorized notion in linguistic theory: not all linguists agree that the notion of Discourse Topic is required in discourse analysis at all (cf. Asher, 2004). For our purposes, however, we formulated a set of patterns for identifying Discourse Topics on the basis of the output of the CMU Link Parser⁶ the system uses.

Paradigmatically, we counted ordinary subjects of the first sentence of a passage as expressive of the Discourse Topic. So, if we found an expression of the topic there, either in full or reduced form, we took that as an instance of the topic appearing as Discourse Topic in that passage

and ranked that passage highly. Of course, not all Discourse Topics are expressed as subjects, and the system recognized this.

A crucial aspect of this functionality is to identify how different sorts of topics can be expressed in a sentence. To give a simple illustration, if the system believes that a topic has been categorized as a personal name, then it accepted reduced forms of the name as expressions of the topic (e.g. "Lindsay" and "Lohan" can both be expressions of the topic "Lindsay Lohan" in certain contexts); but it does not accept reduced forms in all cases.

Paragraphs were verified to contain a sequence of sentences by parsing the rest of the contiguous text. The verb associated with the Discourse Topic of the paragraph was recorded for future use in assembling the topic report. Various filters for length, keyword density, exophoric expressions, spam and obscenity were employed. A score of the intrinsic informativeness of the paragraph was then assigned, making use of such metrics as the length of the paragraph, the number of unique NPs, the type of verb associated with the Discourse Topic, and other factors.

Images were thumbnailed and associated with the extracted paragraph on the basis of matching text in the image filename, alt-text or description elements of the tag as well as the size and proximity of the image to the paragraph at hand. We did not analyze the image itself.

8 Subtopic Selection and Report Assembly

Once the system had an array of extracted paragraphs, ranked by their intrinsic properties, we began constructing the topic report by populating an initial 'overview' portion of the report with some of the best-scoring paragraphs overall.

First, Retriever eliminated duplicate and near-duplicate paragraphs using a spread-activation algorithm.

Next the system applied question-answering methodology to order the remaining paragraphs into a useful overview of the topic: first, we found the best two paragraphs that say *what the topic is*, by finding the best paragraphs where the topic is the Discourse Topic of the paragraph and the associated verb is a copula or copula-like (e.g. *be known as*). Then, in a similar way, we found the best few paragraphs that said *what*

⁶ <http://www.link.cs.cmu.edu/link/>

attributes the topic has. Then, a few paragraphs that said *what the topic does*, followed by a few paragraphs that said *what happens to the topic* (how it is used, things it has undergone, and so on).

The remaining paragraphs were then clustered into subtopics by looking at the most frequent NPs they contain, with two exceptions. First, superstrings of the topic were favored as subtopics in order to discover complex nominals in which the topic appears. Secondly, non-reduced forms of personal names were required as subtopics, even if a reduced form was more frequent.

Similar heuristics were used to order paragraphs within the subtopic sections of the topic report as in the overview section.

Additional constraints were applied to stay within the boundaries of fair use of potentially copyrighted material, limiting the amount of contiguous text from any one source.

Topic reports were set to be refreshed by the system five days after they were generated in order to reflect any new developments.

In an evaluation of 642 paragraphs, 88.8% were relevant to the topic; 83.4% relevant to the topic as categorized. For images, 85.5% of 83 images were relevant, using a revised algorithm, not the live system. Of 1861 subtopic paragraphs, 88.5% of paragraphs were relevant to the assigned topic and subtopic.

9 Discussion

Of the over 30K topical reports generated by Retriever thus far, some of the reports generated turned out surprisingly well, while many turned out poorly. In general, since we paid no attention to temporal ordering of paragraphs, topics that were highly temporal did poorly, since we would typically arrange paragraphs with no regard for event precedence.

There are many things that remained to be done with Retriever, including extracting paragraphs from non-HTML documents, auto-hyperlinking topics within Retriever pages (as in Wikipedia), finding more up-to-date sources for categorization, and verticalizing Retriever page generation for different types of topics (e.g. treating movies differently than people and both differently than diseases). Unfortunately, the project was essentially discontinued in February, 2006.

10 Related Work

Although there have been previous systems that learned to identify and summarize web documents on a particular topic (Allen et al, 1996) without attempting to fuse them into a narrative structure, we are not aware of any project that attempts to generate coherent, narrative topical summaries by *paragraph* extraction and ordering. Much recent work focuses on multi-article summarization of news by *sentence* extraction and ordering (see for example, Columbia's well-known Newsblaster project and Michigan's NewsInEssence project). The latest DUC competition similarly emphasized sentence-level fusion of multi-document summaries from news text (DUC, 2005). One exception is the ArteQuaKt project (Kim et al, 2002), a prototype system for generating artist biographies from extracted passages and facts found on the Web aimed at different levels of readers (e.g. grade school versus university students). The Artequakt system was to use extracted text both as found and as generated from facts in a logical representation. It is not clear how far the ArteQuaKt project progressed.

Less legitimately, more and more "spam blogs" repackage snippets from search results or in other ways appropriate text from original sources into pages they populate with pay-per-click advertising. Retriever differs from such schemes in filtering out low value content and by making obscure sources visible.

References

- Allen, Brad et al. 1996. WebCompass: an agent-based meta-search and metadata discovery tool for the Web. *SIGIR '96*.
- Asher, Nicholas. 2004. Discourse Topic, *Theoretical Linguistics*. 30:2-3
- DUC. 2005 DUC Workshop. Vancouver, BC
- Kim, Sanghee et al. 2002. Artequakt: Generating Talored Biographies from Automatically Annotated Fragments from the Web. In *Proceedings of Workshop on Semantic Authoring, Annotation & Knowledge Markup (SAAKM'02)*. pp. 1-6, Lyon, France.
- Liu, Bing. et al. 2003. [Mining Topic-Specific Concepts and Definitions on the Web](#). Proceedings of the Twelfth International World Wide Web Conference (WWW-2003),
- Shen, Dou et al, Q2C@UST: Our Winning Solution to Query Classification in KDDCUP 2005. *ACM KDD Explorations*. Vol 7, no. 2. December 2005.