

Context-based URL Summarization

Yves Petinot

ypetinot@cs.columbia.edu

Department of Computer Science, Columbia University
503 Computer Science Building
1214 Amsterdam Avenue
New York, New York, 10027
Academic Advisor: Jason Nieh

Abstract

We report on our work on automatic context-based URL summarization. The approach we present here leverages the target URL's context (not only the URL's anchor-text, but also related queries, Web 2.0 tags, etc.) as well as an existing corpus of URL summaries in order to automatically generate short – DMOZ-like – summaries. The generated summaries are intended to supplement the summaries provided by Web directories like DMOZ. While most, if not all, commercial Web search engines rely on DMOZ to provide default URL summaries, DMOZ is manually generated and therefore does not scale in terms of its coverage of the Web. The work presented here aims at providing a viable alternative to DMOZ summaries.

While there has been previous attempts at leveraging URL contexts for summarization [Amitay2000, Delort2003], all approaches have been based on the assumption that summaries, or their component sentences, can be extracted verbatim from the target's context. [Amitay2000] proposed a solution to automatically collect entire summaries from web-pages that link to the site to be summarized. [Delort2003] built up on this idea but, instead of looking for “pre-built” summaries in the target's context, proposed to construct summaries by extracting, clustering and combining sentences from any part of the target's context using sentence topicality and similarity as driving factors.

We argue that, although the context gives a strong indication of the key concepts/topics associated with the target URL, purely extractive approaches are overly restrictive and are unlikely to produce high quality short summaries. Our fusion-based approach [Barzilay2005] combines the target context with existing descriptions of similar URLs in order to generate human-readable summaries that contain the key elements of the context while bearing the characteristics of the summaries of similar URLs.

References

- [Amitay2000]: [Einat Amitay](#), Cécile Paris: Automatically Summarising Web Sites – Is There A Way Around It? [CIKM 2000](#): 173–179
- [Barzilay2005]: Regina Barzilay, [Kathleen McKeown](#): Sentence Fusion for Multidocument News Summarization. [Computational Linguistics 31](#)(3): 297–328 (2005)
- [Delort2003]: Jean-Yves Delort, [Bernadette Bouchon-Meunier](#), [Maria Rifqi](#): Enhanced web document summarization using hyperlinks. [Hypertext 2003](#): 208–215
- [Sun2005]: [Jian-Tao Sun](#), [Dou Shen](#), [Hua-Jun Zeng](#), [Qiang Yang](#), [Yuchang Lu](#), Zheng Chen: Web-page summarization using clickthrough data. [SIGIR 2005](#): 194–201