Site Search Using Profile-Based Document Summarisation

Azhar Alhindi
University of Essex,
Colchester, UK
ahalhi@essex.ac.uk

Udo Kruschwitz
University of Essex,
Colchester, UK
udo@essex.ac.uk

Chris Fox
University of Essex,
Colchester, UK
foxcj@essex.ac.uk

ABSTRACT

Text summarisation is the process of distilling the most important information from a source to produce an abridged version for a particular user or task. This demo presents the use of profile-based summarisation to provide contextualisation and interactive support for site search and enterprise search. We employ log analysis to acquire continuously updated profiles to provide profile-based summarisations of search results. These profiles could be capturing an individual's interests or those of a group of users. Here we look at acquiring profiles for groups of users.

1. MOTIVATION

Summarisation is a broad area of research [8]. The sort of information contained in a summary differs according to the mechanism used in the summarisation process: It may highlight the basic idea (generic summarisation), or it may highlight the specific user's individual area of interest (personalised summarisation). One of the techniques used to achieve personalisation is user profiling. User profiles may include the preferences or interests of a single user or a group of users and may also include demographic information [4]. Normally, a user profile contains topics of interest to that single user. We are interested in capturing profiles not of single but groups of users.

We utilise query and click logs to acquire a profile reflecting the population's search patterns and this profile is being automatically updated in a continuous learning cycle. We are then applying the acquired profiles in the summarisation process to support users searching a document collection. The potential of personalised summarisation over generic summaries has already been demonstrated, e.g. [3], but summarisation of Web documents is typically based on the query rather than a full profile, e.g. [11, 9]. Our specific interest lies in enterprise search which is different from Web search and has attracted less attention [5]. The benefit of this context is that we can expect a more homogeneous population of searchers who are likely to share interests and information needs. Our hypothesis is that profile-based summarisation can help a user in this process and guide the user to the right documents more easily (e.g. by presenting the summaries instead of or alongside snippets).

2. METHODS AND EXAMPLES

The demo presents an integrated Solr-based search system applying a number of different methods for building summaries for search results. The first two algorithms were designed for traditional (generic) summarisation, and they represent widely used baselines, e.g. [12]. The other three are all variations of an approach that has been proposed in the literature for building an adaptive community profile/domain model, a *"biologically inspired model based on ant colony optimisation applied to query logs as an adaptive learning process"* [1]. The approach is simple to implement, the idea here is that query logs are segmented into sessions and then turned into a graph structure. Figure 1 gives an example of part of the profile as it has been derived from our query logs. We used the log files collected on the existing search engine over a period of three years[1] to bootstrap this ant colony optimisation (ACO) model, i.e. our profile. The example illustrates the domain-specific nature of the derived profiles, e.g. the University library is named after Albert Sloman.

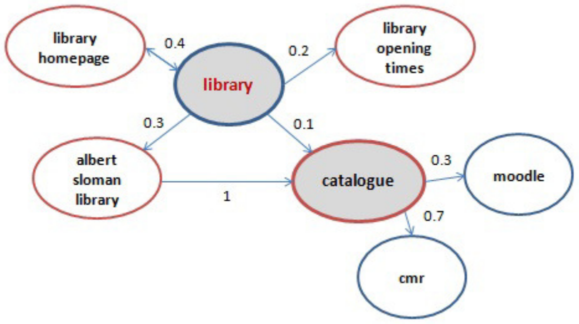

Figure 1: Partial profile derived from query logs.

A profile-based (extractive) summary of a document is then generated by turning the profile into a flat list of terms (we use three different methods to do this as explained further down) and selecting those sentences from the document

DIR 2013, April 26, 2013, Delft, The Netherlands.
.

[1] More than 1.5 million queries, described in more detail elsewhere [6]

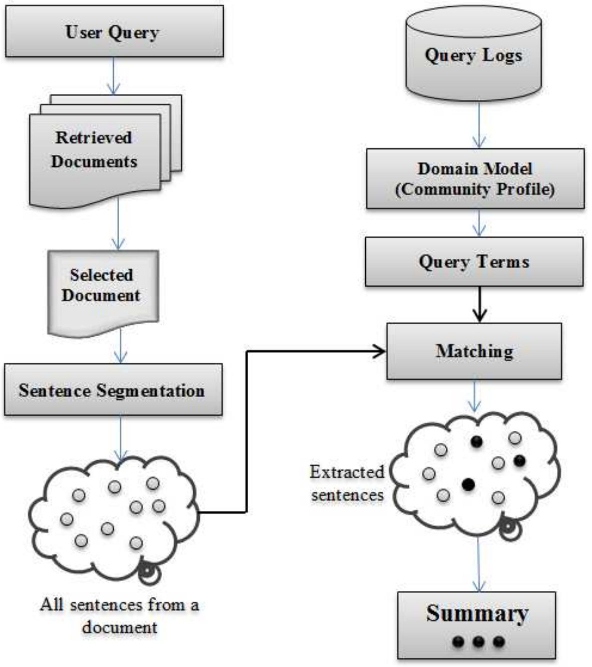

Figure 2: Architecture of profile-based single-document summariser.

that are most similar to the profile using cosine similarity. Figure 2 shows an architectural diagram for our profile-based summarisation system. Following DUC 2002 convention we select 100-word abstracts [7]. This gives us the following five methods:

1. **Random:** Selects sentences from the document randomly [12].
2. **Centroid:** A centroid-based approach to summarisation [10]. This algorithm takes into account first-sentence overlap and positional value, which are then used to generate a coherent summary.
3. **ACO:** A query graph built by processing the log data according to [1]. The entire model is turned into a flat list of terms for summarisation.
4. **ACO trimmed:** Starting with *ACO* we trim all those edges whose weights fall below the overall average weight of an edge. The remaining model is turned into a flat list of terms for summarisation.
5. **ACO query refinements:** The list of terms used for summarisation are all those that are directly linked to the query node in the ACO model.

Note that **ACO** and **ACO trimmed** are query-independent as they are using the *entire* model to generate the summary, whereas **ACO query refinements** is query-specific (for the frequently submitted query *"library"* the selected terms are *library*, *albert sloman library*, *library homepage*, *library opening times* and *catalogue*, see Figure 1).

To illustrate the different summaries obtained using three of the summarisation methods we apply the methods to the University of Essex Library homepage[2] and get the following summaries:

- **Random:** *We provide you with first class library facilities to complement and assist your studies. The Library allows you to access more than one million books, periodicals and microfilms. Our collections covering Latin America, Russia and Eastern Europe are of national significance. Viewing facilities for DVDs and videos are also available.*

- **Centroid:** *We provide you with first class library facilities to complement and assist your studies. You can find out more by accessing our University library services. Our Albert Sloman Library is just a few minutes' walk from teaching buildings and student accommodation at our Essex Campus.*

- **ACO query refinements:** *In addition, 110 networked PCs and terminals provide access to over 47,000 online journals, databases, e-books and library catalogues. Students at our Essex Campus can visit the Albert Sloman Library or borrow books from its collection via a daily dispatch service. The Albert Sloman Library has long opening hours, a total of 84 hours over seven days a week during term and 42.5 to 84 hours in vacations.*

Obviously, the actual usefulness of such summaries can only be assessed in a realistic search setting. In a pilot study we found that the ACO-based summaries have the potential of outperforming the different baselines [2]. A task-based evaluation using TREC Interactive Track guidelines is currently being conducted. As the immediate next step, we are interested in investigating how the profile can be integrated into multi-document summarisation.

3. REFERENCES

[1] M-D. Albakour, U. Kruschwitz, N. Nanas, D. Song, M. Fasli, and A. De Roeck. Exploring ant colony optimisation for adaptive interactive search. In *Proceedings of ICTIR*, pages 213–224. Springer, 2011.

[2] A. Alhindi, U. Kruschwitz, and C. Fox. A pilot study on using profile-based summarisation for interactive search assistance. In *Proceedings of ECIR*, pages 672–675, 2013.

[3] A. Díaz and P. Gervás. User-model based personalized summarization. *Information Processing & Management*, 43(6):1715–1734, 2007.

[4] S. Gauch, M. Speretta, A. Chandramouli, and A. Micarelli. User profiles for personalized information access. *The Adaptive Web*, pages 54–89, 2007.

[5] D. Hawking. Enterprise Search. In R. Baeza-Yates and B. Ribeiro-Neto, editors, *Modern Information Retrieval*, pages 641–683. Addison-Wesley, 2nd edition, 2011.

[6] U. Kruschwitz, D. Lungley, M-D. Albakour, and D. Song. Deriving Query Suggestions for Site Search. *JASIST*, 2013. Forthcoming.

[7] C.Y. Lin and E. Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of HLT-NAACL*, pages 71–78. ACL, 2003.

[8] A. Nenkova and K. McKeown. *Automatic summarization.* Now Publishers, 2011.

[9] S. Park. Personalized summarization agent using non-negative matrix factorization. *PRICAI 2008: Trends in Artificial Intelligence*, pages 1034–1038, 2008.

[10] D.R. Radev, H. Jing, M. Stys, and D. Tam. Centroid-based summarization of multiple documents. *Information Processing & Management*, 40(6):919–938, 2004.

[11] C. Wang, F. Jing, L. Zhang, and H.J. Zhang. Learning query-biased web page summarization. In *Proceedings of CIKM*, 2007.

[12] R. Yan, J.Y. Nie, and X. Li. Summarize what you are interested in: An optimization framework for interactive personalized summarization. In *In Proceedings of EMNLP*, pages 1342–1351, 2011.

[2] http://www.essex.ac.uk/life/library/