EXAMPLE-BASED SEARCH FOR TOURIST INFORMATION

Line Eikvil and Kjersti Aas
Norwegian Computing Center
P.O.Box 114 Blindern, 0314 Oslo, Norway
TEL: +47 22 85 25 00, FAX: +47 22 69 76 60

E-mail: Line.Eikvil@nr.no, Kjersti.Aas@nr.no

Abstract
Search tools are an important feature of a tourist information system. Traditionally search facilities have been based on keywords or limited database fields. Tourists may however often want to illustrate what they are interested in by giving an example. Hence, methods for example-based search would be a useful addition to such systems. This paper describes an example-based search facility developed for a Web-based booking system. The search facility is based on statistical methodology, and uses a concept-based method that assumes the existence of some underlying semantic structure in the tourist information data. The methods have proved useful for a number of applications and seem to be a useful addition to traditional methods also for a tourist information system.
Key words: tourist information system, search facilities, example-based search.

1 Introduction

This paper describes a search tool developed for a tourist information system. Search tools are an important feature of such a system. In (Tjøstheim, 2000) a survey has been performed among Norwegian tourists and Internet users concerning the importance of different functionalities in an Internet based service for tourist information. The survey has been performed amongst two groups of people, where the first group was selected from all people having access to Internet, and the second group from those that have access to the Internet and also plan to book a holiday through the Internet.

One of the questions asked in this study, was to rate the importance of good search functionalities on an Internet tourist service. Among the first group 51% meant that search functionality was very important, while 24% meant that this was quite important. In the second group 68% said that search functionality was very important, while 22% said this was quite important. Compared with other functionalities, search scored very high for both groups.

Traditionally search facilities have been based on keywords or limited database fields, and these are important functionalities in a system. However, this does not necessarily accommodate the true diversity of searchers needs, and can make it difficult for the searcher to locate approximate information. This is bad for the searcher and for the information provider.

In (Bullock and Goble, 2000) the findings from an ethnographic study performed in tourist information centres to determine how tourists ask for information, and how tourist advisers respond, are reported. These studies revealed that tourists are often unsure of what is available and do not have specific places in mind. Instead they describe what they would like and want the adviser to suggest suitable places. Their descriptions vary in form and clarity from quite simple to complex descriptions where the customer paints a narrative picture of what they would like. Traditional search methods are not able to accommodate these types of queries, and more intelligent search methods are needed.

2 Related work

Standard information-retrieval methods depend on exact matches between words in users' queries and words in documents. They usually treat words as if they are independent, although it is quite obvious that they are not. This exact word match also fails because there are many ways to refer to the same object, and many words have more than one distinct meaning. In addition, it has been found that for a single well-known object, two people will use the same main key word less than 20% of the time. Due to these problems, more intelligent search methods have appeared. There are three major information retrieval paradigms: *contextual*, *semantic* and *statistical*, and in what follows, a review of some of the work that has been done in these areas is given.

The first approach takes advantage of the structural and *contextual* information typically available in retrieval systems. For example, this could involve the use of a thesaurus and encoded relationships among terms. Thus when the search terms are specified, the search tool also looks for the related words. The thesaurus is often constructed manually, but may also be created automatically. The latter can be done by determining words having high probability of occurring together in the documents in the database. Due to the difficult and time-intensive nature of creating a thesaurus, this method is applicable only to small databases, which are relatively narrow in their subject matter. Automatic construction is less time-consuming than manual, but the search precision tends to be severely affected.

The *semantic* approach to information retrieval views documents and queries as representing some underlying meaning. It emphasizes natural language processing or the use of artificial intelligence. The concepts underlying this field are rooted in theoretical linguistics, and is based on understanding of lexicons, grammars, syntax and semantics. A problem with natural language processing techniques is that they are generally not designed to handle large amounts of text from many different domains, and specific knowledge of the application domain is usually required.

The third approach emphasizes *statistical* correlations of word counts in documents and document collections. Using statistical methods relationships between words are incorporated indirectly using statistical algorithms. From these methods it is also straightforward to create a similarity score between any two documents based entirely on the words that they contain. One example of a statistically based search tool is the so-called Intelligent Concept Extraction of Excite (Excite). Another is the Frequently Asked Question Organizer (FAQO) described in (Caron, 2000). In the FAQO prototype, an approach denoted Latent Semantic Analysis (LSA) is used for query matching. LSA uses statistical correlation between words within a body of documents to infer underlying semantic concepts between both the documents and the words. This means that if word correlations have been correctly extracted, two documents may be found to be quite similar even if they have no words in common, thus giving this technique clear advantage over simple term-matching information retrieval methods.

For the tourist information search, a statistically based method has been chosen. There are three main reasons for this. First, the database is not sufficiently narrow in its subject matter to make thesaurus-based methods or methods based on natural language processing practical. Second, statistical methods are better suited for example-based search. Finally, these methods will facilitate a ranking of the search results. The statistical method chosen for this application is based on singular value decomposition and shares many characteristics with the method described in (Caron, 2000).

Such methods have also been used in various other information retrieval applications, including digital library text retrieval (Dumais, 1995), cross-language document retrieval (Dumais et al., 1997), automatic grading of papers (Rehder et al., 1998), document classification (Laham, 1997), and image retrieval (Penenovic, 1997). Hence, the application area for these methods seems to be wide.

3 System

The example-based search tools described in this paper will be part of a Web-based booking system for the Norwegian tourism industry, which is developed by DBC-Munin, (Norbooking, 2000). Figure 1 shows an example of one of the Internet search pages of this service. The purpose of this system is to make it possible

to plan holidays and make travel reservations directly on the Web. An important functionality of such a system is the search capabilities. In DBC-Munins' system the user will be offered to search for information in four different ways:

- *navigate* through the information using hypertext-links
- perform a *structured search* on the basis of a common database schema
- perform a *keyword-based search*
- perform an *example-based search*

This Web-based tourist information service aims at covering the whole of Norway, offering booking of smaller lodgings like cottages, apartments and guesthouses as well as different activities. The data in this information system will consist of descriptions of the lodgings and activities, where the data used for the example-based search consists of:

- a description of the particular product, e.g. the cottage, the motel, an activity etc.
- a description of the provider of this product, containing general information about the area.
- a list of available facilities like TV, sauna, swimming pool, fireplace etc.
- a short description of the site, e.g. by the seaside, in the mountains, etc.

All this textual information is combined to give the description of each product contained in the database. The conceptual search is then performed based on this combined description.

Figure 1: *Screen from Norbooking.*

4 Methods

The example-based search is performed on textual data exported from the main database. The textual data are organised in a representation that enables concept-based searches. This representation is generated by first transforming the collection of texts into a numerical matrix, and then performing a singular value decomposition (SVD) on this matrix. Singular value decomposition is a technique closely related to eigenvector decomposition and factor analysis, and it is used here to model the associative relationships between documents. The first step is treated in Section 4.1, while the latter is described in Section 4.2.

Assuming that an SVD-representation exists, an example-based query is handled by first transforming the query text to a numerical vector, and then comparing this vector to the SVD-representation to find the best match(es). This issue is treated in Section 4.3, while Section 4.4 describes how the example-based search tools are integrated into the total system.

4.1 Document representation

The first step in automatic analysis of text documents is to transform the documents into a representation suitable for the analysis. A commonly used document representation is the so-called vector space model (Salton, 1983). In the vector space model, documents are represented by vectors of words. Usually, one has a collection of documents which is represented by a word-by-document matrix **A**, where each entry represents the occurrences of a word in a document, i.e.:

$$\mathbf{A} = (a_{ik}) \tag{1}$$

where a_{ik} is 1 if word i occurs in document k and 0 otherwise.

Words that do not carry much information, e.g. prepositions and pronouns, are removed from the word set. Such words are often denoted *stopwords* in text analysis. Then, each word is reduced to its *stem* to get the same representation for inflected forms.

4.2 SVD analysis

SVD analysis is a concept-based method that assumes the existence of some underlying semantic structure in the text data matrix. It further assumes that this structure is partially obscured by the randomness of word choice, but that it can be estimated by statistical techniques. Dependencies between words are explicitly taken into account in the representation by simultaneously modelling all the interrelationships among terms and documents. In what follows, an outline of the method is given.

Assume that there exists an MxN word-by-document matrix **A**, where M is the number of words, and N the number of documents. The singular value decomposition of **A** is given by:

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{2}$$

where **U** (MxR) and **V** (RxN) have orthonormal columns and $\mathbf{\Sigma}$ (RxR) is the diagonal matrix of singular values. R <= min(M, N) is the rank of **A**. If the singular values of $\mathbf{\Sigma}$ are ordered by size, the K largest may be kept and the remaining smaller ones set to zero. The product of the resulting matrices is a matrix $\mathbf{A_K}$ which is an approximation to **A** with rank K:

$$\mathbf{A}_K = \mathbf{U}_K\mathbf{\Sigma}_K\mathbf{V}_K^T \tag{3}$$

Here $\mathbf{\Sigma}_K$ (KxK) is obtained by deleting the zero rows and columns of $\mathbf{\Sigma}$, and $\mathbf{U}_K$ (MxK), and $\mathbf{V}_K$ (NxK) are obtained by deleting the corresponding rows and columns of **U** and **V**.

The idea is that $\mathbf{A}_K$, by containing only the first K independent linear components of **A**, captures the major correlation structure in the matrix and removes the noise. Since K will usually be much smaller than the number of words M, minor differences in terminology are ignored. The similarity of documents is determined by the overall pattern of word usage, so that documents may turn out to be similar even if they have no words in common.

4.3 Performing queries

When a word-document matrix has been generated, and the SVD of this matrix has been computed, a query to this SVD can be performed.

First the query is parsed and a document vector **d** is generated in the same way as described in Section 4.1. For **d** to be compared with the documents represented in the SVD database, it must be represented as a vector in the same K-dimensional space. This is achieved by computing the projection:

$$\hat{\mathbf{d}} = \mathbf{d}^T\mathbf{U}_K \tag{4}$$

The cosine between $\hat{\mathbf{d}}$ and the rows of $\mathbf{V}\mathbf{\Sigma}_K$ gives the degree of similarity between the new document and the documents in the database. The cosine will be on the interval [-1,1]. A cosine of 1 means that the two are identical with respect to the chosen representation, while a similarity of -1 means that they have nothing in

common. The documents in the database are sorted according to the similarity with the query before they are returned.

4.4 Integration of the methods in the system

The SVD-analysis described in Section 4.2 can be a rather time-consuming process. However, this analysis need only be performed when there have been changes to the database, and there is no need to do this analysis online. Hence, when necessary these computations can be performed offline independent of the rest of the system. Then when a new SVD has been computed, the online-system can be updated. The SVD analysis is therefore not an integrated part of the online system, but runs as a separate stand-alone program.

The matching process, where the query posed by the user is matched against the SVD matrix, must of course be performed online. The actual matching process as described in Section 4.3, is however quite fast once the SVD matrix has been read into memory. This online query phase is implemented in Java and integrated with the rest of the system.

Example-based search is not perfect for every type of search. For very explicit queries, specifying for instance the number of beds in a cottage, or the availability between specified dates, a traditional database search is usually required. Therefore, a combination with database search is planned, where the example-based search can be performed on a specified set of documents obtained through an initial database search (e.g. for availability during a specified period), instead of the entire database.

5 Experiments and results

The development of the search system was performed early in the process of the development of the system. This meant that the amount of data available in the database was very limited. A certain minimum size of the database is, however, required for the statistically based search methods to show their potential. The example-based search methods described in this report have therefore not yet been made available to the public, and the experiments and results reported in this chapter are based on some initial tests.

5.1 Data

The data exported from the database are ordered by product, where a product is defined as a specific cottage, hotel, activity etc. Each product has a textual description, which is combined with information on the available facilities, the site and a more general description given of the provider of the lodging or activity. The product description is then specific to each product, while the provider description could be the same for several products. The average length of the texts is approximately 115 words per product, and all the texts are in Norwegian.

As the experiments reported here were performed early in the process of the development of the system, the amount of available data in the tourist database was limited. Hence, only about 70 different product descriptions were available, which is a very small database. There are two main problems with performing experiments with the statistical search method on such a small database:

- The variety of the offers in the database is very limited. Hence, there is a high probability that a potential user will ask for something which is not contained in the database.
- The example-based search method performs an analysis of the associations among words and documents, and produce a representation in which words that are used in similar contexts will be more semantically associated. However, with very few documents, few associations can be captured from the data, and this will affect the performance.

Anyhow, some tests on these data were performed to get at least an initial indication of what can be expected from the search method for this application.

5.2 Experiment

To get an evaluation of the results, a few persons tested the search on the described database. They were asked to formulate queries in natural language describing what they would like to do and the type of place they would like to visit for a holiday. Each person was asked to formulate a few different queries.

For each query in this experiment the five highest ranked results were returned, and the descriptions of each of these five sites were displayed to the user. The test persons were then instructed to evaluate each of the returned documents according to the aim of their query. They were asked to give a response of either **Yes**, **Maybe** or **No** to whether they found that the returned document matched their query.

As the database was quite small, the users were also given an idea about what was in the database, and what area of Norway the database covered. However, the contents could of course not be described in detail, and the probability that some of the test persons would ask for something that was not in the database was quite high.

5.3 Performance

As expected the results of this small study were influenced by the problems mentioned in Section 5.1. The main problem seemed to be that although the users were given an idea about what was in the database, the test persons often asked for something that was not contained in the database. It therefore became very difficult to do a quantitative analysis of the results. The overall impression was, however, that people asking for something which was present in the database were content with the returned answers to their query.

The example-based method is supposed to overcome problems in retrieval due to different word usage. With a small database like this, there are however, not enough examples to enable the necessary associations between words. Hence, for users requesting e.g. less frequent activities, the words had not appeared sufficiently often for the method to make the necessary associations.

When enough data have been collected in the database, the relevance of the returned results should be tested more thoroughly. At this point the methods do however seem to give relevant answers when relevant product descriptions are available.

6 Discussion and conclusion

The preliminary tests of performance of the methods, indicate that they can be well suited for example-based queries to a tourist database. However, for small databases like the one in our experiments, it is difficult to evaluate the method's true potential. There is a high risk of the queried elements not being in the database, and there are too few examples to make enough associations.

There is a need to test the system on a larger database to get a true picture of its performance. A test should therefore be performed when the system has been running for a while, to get the users' opinion of the search facility. This search functionality does however seem to be a useful addition to the more traditional search facilities.

In addition to allowing for example-based search, the method described in this paper can also provide other functionalities. These techniques can for instance be used to make recommendations to the user by finding documents similar to their original choice. This can be helpful for instance when the chosen item is not available for the desired period. Recommendations can even be made based on choices of other users with similar queries. The methods have also been applied for cross-language retrieval, which is another functionality that could be useful for the current application. In conclusion it seems that this technology will offer new and interesting possibilities to a tourist information system.

References

Norbooking. http:/www.norbooking.no/

Intelligent Concept Extraction from Excite. http://www.excite.com/ice/tech.html

Excalibur Solutions. http://www.excalib.com/solutions/internet.shtml

Bullock, J.C. and Goble, C.A. (2000). Putting the tourist into Tourist Information, *Proceedings of ENTER 2000, Barcelona*, Springer Wien-New York, pp.104-113.

Caron, J. (2000). Applying LSA to Online Customer Support: A Trial Study. April 2000.

Dumais, S.T. (1995). Using LSI for information filtering: TREC-3 experiments. *In D. Harman (Ed.), The third Text Retrieval Conferenec (TREC3)*, National Institute of Standards and Technology Special Publication.

Dumais, S.T., Letsche, T.A., Littman, M.L. and Landauer, T.K. (1997). Automatic cross-language retrieval using Latent Semantic Indexing. *In AAAI Spring Symposium on Cross-Language Text and Speech Retrieval*, March 1997.

Laham, D. (1997). Latent semantic analysis approaches to categorization. *In: M.G. Shafto and M.K. Johnson (eds.), Proceedings of the 19th Annual Conference of the Cognitive Science Society*, Mahwah, NJ:Erlbaum, p. 979.

Pecenovic, Z. (1997). Image retrieval using Latent Semantic Indexing. *Diploma Thesis, Ecole Polytechnique Fédérale de Lausanne*.

Rehder, B., Schreiner, M.E., Wolfe, M.B., Laham, D., Landauer, T.K. and Kintsch, W. (1998). Using Latent Semantic Indexing Analysis to assess knowledge: Some technical considerations. *Discourse Processes 25*, pp. 337-354.

Salton, G. and McGill, M.J. (1983). An Introduction to Modern Information Retrieval, *McGraw-Hill*, 1983.

Tjøstheim, I. (2000). Travel sites and the online holiday market - a national study. *NR - Report*, December 2000.