# Web Usage Mining in Tourism – A Query Term Analysis and Clustering Approach

Arthur Pitman[a], Markus Zanker[a],
Matthias Fuchs[b/c] and
Maria Lexhagen[b]

[a] Universität Klagenfurt, Austria
{arthur.pitman,markus.zanker}@uni-klu.ac.at

[b] European Tourism Research Institute (ETOUR), Mid Sweden University, Sweden
{matthias.fuchs, maria.lexhagen}@etour.se

[c]eTourism Competence Center Austria (ECCA),
University of Innsbruck, Austria
{matthias.fuchs}@ecca.at

## Abstract

According to current research, one of the most promising applications for web usage mining (WUM) is in identifying homogenous user subgroups (Liu, 2008). This paper presents a prototypical workflow and tools for analyzing user sessions to extract business intelligence hidden in web log data. By considering a leading Swedish destination gateway, we demonstrate how query term analysis in combination with session clustering can be utilized to effectively explore the information needs of website users. The system thus overcomes many of the limitations of typical web site analysis tools that only offer general statistics and ignore the opportunities offered by unsupervised learning techniques.

**Keywords:** Web usage mining, query term analysis, clustering, destination portal

## 1 Introduction

Knowledge about the users of a web site, their information needs and search behavior is crucial for ensuring the effectiveness of online marketing (Biswas & Krishan, 2004; Dias & Vermunt, 2007). A web server's log file is one commonly available data source for learning about visitors' information needs, however, it is often left unexploited. Standard analytics tools (e.g. *Google Analytics* or *123loganalyzer*) provide solely descriptive information about page access frequencies, view times, common entry and exit points, referral sites, etc. and thus provide a blurred picture of online behavior. In contrast, web mining aims to discover useful knowledge (i.e. business intelligence) from the structure of hyperlinks (i.e. web structure mining), page content (i.e. web content mining) and usage data (i.e. web usage mining) (Liu, 2008, p. 6). Statistical techniques such as *session analysis* may be applied capture the frequency of page accesses and search words, path lengths, entry and exit points as well as referral sites. Subsequently, the mining methods most frequently applied in web usage mining (WUM) are unsupervised learning techniques and rule mining

approaches. *Unsupervised learning* techniques, such as clustering, can be used to discover both user clusters (of sessions or transactions) as well as page clusters (Larose, 2005). This is especially useful when analyzing market segmentation in e-commerce or to provide personalized web content for users with similar interests (Bhatnagar & Ghose, 2004). *Association rule mining* may be used to find groups of items or pages that are commonly accessed or purchased together, thus, organize content more effectively or to cross-sell product items (Mobasher, 2008, p. 471). Finally, s*equential pattern mining* can be applied to capture frequent navigational paths among user trails. In this paper we propose a WUM approach that combines query (i.e. search) term analysis and user clustering. The proposed approach is then empirically tested in the context of a leading online portal of the tourism destination Åre, Sweden. The paper is structured as follows: in Section 2 we examine related work in the area of WUM in tourism. Following this, Section 3 outlines our proposed data mining workflow, including various post processing steps. Section 4 presents results from the analysis of search terms used in referrals and on the site itself as well as results from clustering to identify homogenous web-user groups. The conclusion summarizes the managerial implications and proposed future research.

## 2 Related Work

Although tourism is dominated by e-business systems and applications (Werthner & Ricci, 2004), to date relatively few attempts have been made to systematically explore the huge potentials of WUM in the e-tourism domain. Some of the few exceptions include: Tichler et al. 1999; Murphy et al. 2001; Cho & Leung, 2002; Olmeda & Sheldon, 2002 and Honda et al. 2006. For the remainder of this section we discuss two recent examples of tourism related WUM. In their attempt to model the navigation behavior of hotel guests, Schegg et al. (2005) analyzed log-files from 15 Swiss hotels. Their findings show that an average visitor stays almost 2 minutes at a site and views 4.7 pages, with the most requested pages being the homepage, information related pages (e.g. hotel information, room information) and transactional pages (e.g. booking, guest book). In half of the cases no referring site was registered, which implies that the visitors typed the domain name directly into their browser, clicked a bookmark, favorite or link in an email, thus suggesting a high degree of website familiarity among users. The authors also identified the top 10 search words (e.g. hotel name, location and tourism activities) and referring search engines as well as tourism websites (e.g. online intermediaries, destination websites, etc.). Similarly, Wolk and Wöber (2009) extracted search words from log files generated by the domain specific search engine 'European Cities Tourism' which is comprised of 186 European touristic cities (www.visiteurope.info). The results revealed 5,550 different search words. Interestingly, the top 100 most frequent search words covered over 75% of all search queries. After transforming the data into a table with the percentages of a specific search word (i.e. columns) and the cities (i.e. rows), search profiles were deduced for a subset of 32 cities, serving as the input for strategic positioning analysis.

## 3 Methodology

Log files do not permit easy analysis: their sheer volume and simple structure make it difficult to extract business intelligence directly (Liu, 2008). Thus, before applying the unsupervised learning techniques presented in this study we utilized an in-house tool known as *Web Log Analyzer* (WLA) to reconstruct individual user sessions from the raw weblog data and store them in a standard relational database. Its functionality is somewhat equivalent to that described in Mobasher (2008). The advantage of storing extracted data in a database becomes evident when exploring opportunities for further data processing. Business intelligence may be extracted directly from the database using SQL queries or by applying WUM algorithms. In this project we conducted our initial analysis using a series of SQL queries and then further explored the data using *RapidMiner* (http://www.rapid-i.com) and *R* (http://www.r-project.org), two open source packages specializing in data mining and statistics respectively.

The website *www.visitare.se* is a leading tourism portal for the region of Åre, Sweden. It specializes in offering information about skiing, restaurants, accommodation, sightseeing and services, and is available in Swedish, English and German. Importantly, the site does not support bookings or purchases itself, but rather functions as a referrer to partner sites like *Holiday Club* and *SkiStar*, meaning that no obvious conversion ratio is available. Its category-based design is, however, ideal for WUM. The evaluation, illustrated in Figure 1, comprised of examining log data collected from the gateway over an eight month period between August 2008 and March 2009.

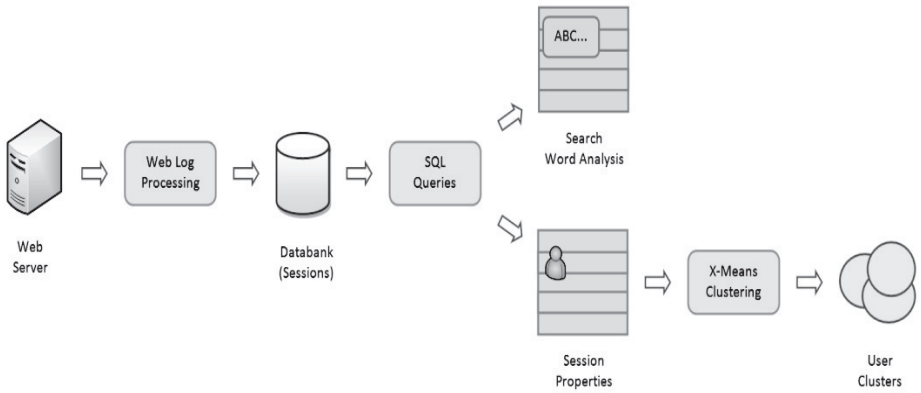


**Fig. 1.** The web-usage mining workflow

In a first step WLA was used to extract sessions from the raw log files. We specified that requests for images and cascading style sheets should be ignored, focusing instead on sequences of GET and POST requests for HTML files. In addition, we applied patterns to mark key user actions within the sessions, such as search activity, requests for PDF brochures and category-based browsing, as well as ambient information, such as interface language changes. Moreover, of great interest was also

how the user entered the site (i.e. through referral from another site or directly via a bookmark or entering the address).

Following this, a series of SQL queries were applied to aggregate this information, producing 106 variables for each session describing temporal aspects (e.g. session duration, start time and day), how the user was referred to the site (e.g. whether the session was the result of a Google search and if so the used search terms and categories) and which areas were viewed (e.g. information categories and subpages), as well as use of special functionality (e.g. search forms or PDF brochures). In addition, we exported search word profiles for Google referrals and the site's internal text search.

We hypothesized that differences in topical variables, such as the categories a user visited or searched for, would be correlated with changes in other variables, particularly in those viewed as success criterion, such as session duration or the number of PDF documents.

## 4 Findings

The web log analysis revealed a total of 183,728 sessions over the eight month period in which web log data was collected. Of these, 63,648 could be attributed to bots and web crawlers, and a further 28,045 sessions contained negligible activity (i.e. failed to request an actual page) and were thus excluded. Of the remaining 92,035 sessions, 33,981 were referred to the site from Google, the result of searches involving an average of three terms, a figure that fully supports the findings of Orlando and Silvestri (2009). In addition, 13,174 sessions searched the site internally using its internal text search facility with an average of 1.6 words per query and 7,092 used the parametric search form.

Search Term Analysis

The processing of search words presents a number of challenges. Many words may be written in a number of different ways, using multiple spelling, declination or punctuation patterns (Honda et al. 2006; Wolk & Wöber, 2009). In our analysis, we treated search terms in a case insensitive fashion, removed connector words (such as "and" in English and "i" in Swedish), merged multiple spellings or declinations and removed punctuation.

When comparing external search terms used in Google searches to those used in internal searches, it turned out that internal searches are much more specific, usually referring to particular categories or items of interest. Clearly, the top Google referral keyword is the name of the region itself, *Åre*, as it is the key term that differentiates the site from others. Other non-Swedish search terms presumably direct users to other sites, thus, do not appear in the log as frequently. The distribution of referral search terms also basically replicates the results of Wolk and Wöber (2009), with the first 150 terms covering more than three quarters of the search queries (see concentration

curve in Figure 2). In contrast, internal searches exhibit much greater diversity with around 250 search terms covering about 75% of all queries.

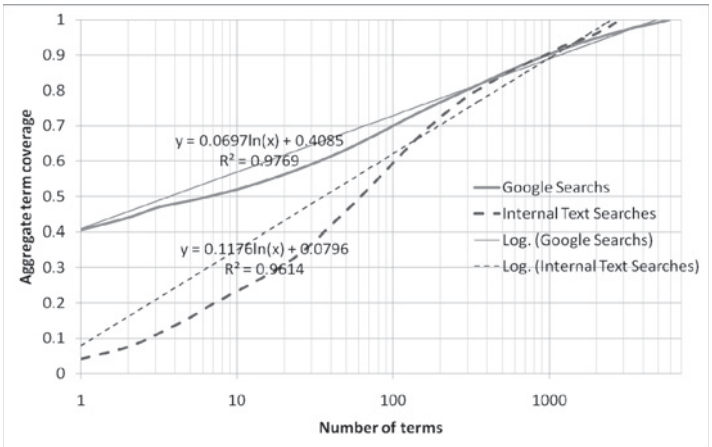


**Fig. 2.** Coverage of Google and internal search terms

Follow this initial analysis, 290 of the most common search terms were assigned to the eight categories by domain experts with the aim of deriving semantic information to better understand the website's various user groups. The resulting distribution can be summarized as follows: *accommodation* (12%), *activities* (16%), *skiing* (7%), *dining* (7%), *shopping* (3%), *attractions* (7%) and *services* (17%), together with *non-specific* terms (32%).

User Cluster Analysis

Perhaps even more interesting than the general behavior of the site's users is the behavior of homogenous sub groups (Dias & Vermunt, 2007). We explored cluster analysis, a technique which assigns items to subsets according to a similarity criterion to understand how the *VisitAre* gateway serves the interests of different types of visitors to the region. Broadly speaking, *x-means* clustering was applied to a subset variables preselected using *principal component analysis* (PCA). Following this we explored statistically significant differences between the resulting clusters in other variables.

*X-means*, a variation of *k-means* that determines *k* by optimizing an effectiveness criterion, avoids the difficulty of having to fix the number of clusters beforehand (Pelleg & Moore, 2000; Hastie et al. 2009). PCA, usually employed as a method for transforming potentially correlated variables into a series of uncorrelated components (Jolliffe, 2002), was used to identify which variables were responsible for the majority of variance thus reducing the amount of explicit domain knowledge required.

Seven variables were selected by PCA as clustering input variables, namely category browse actions for *'accommodation'*, '*to do'*, '*to see'*, '*dining'*, '*service'*, '*communications'* and '*program*'. *X-means* clustering carried out for $2 \leq k \leq 30$ revealed four stable clusters. As can be seen in a decision tree constructed using the clustered data, the focus of the clustering was on *accommodation* (Figure 3). A more detailed summary of the properties of each cluster is presented in Table 1.

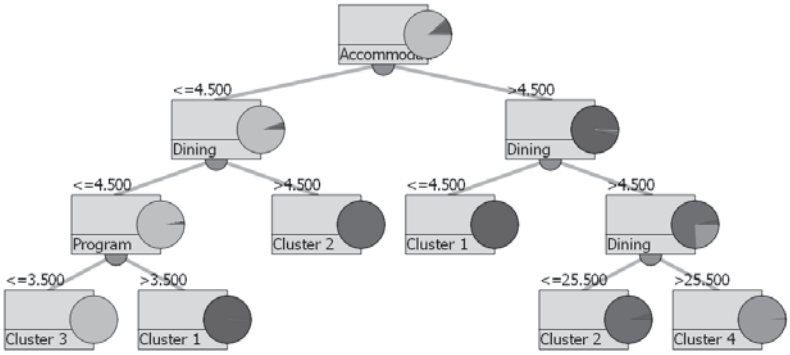


**Fig. 3.** Decision tree constructed using the clustered data

Cluster 1 contained sessions that were primarily interested in *accommodation* and to some extent *program*. Cluster 2, on the other hand, was made up of sessions that concentrated on *dining* whilst also being interested in most other categories. Sessions in cluster 3, the largest cluster, were focused on *accommodation* and on *to do* to some extent, however were on average significantly shorter than those in cluster 1 and cluster 2, and, with the exception of viewing PDF files, visited fewer areas of the site.

Without additional information it is difficult to pinpoint the cause for this behavior. It is possible, for example that sessions in cluster 3 (i.e. the majority of users) were not particularly interested in any single category, navigated to site in error or found it irrelevant, or even found the required information immediately. On the other hand, sessions in cluster 1 and 2 found the offerings much more useful or encouraging, browsing in specific categories and remaining on the website for significantly longer periods of time, thus, showing a more focused and effective search behavior (Dias & Vermunt, 2007). Sessions in cluster 1 and 2 were also significantly more likely to enter the site through the main page rather than one of the specific pages, presumably then navigating to pages of interest.

Finally, cluster 4, a very small cluster, contained sessions that appeared to have been abnormally interested in all categories. Despite using neither the internal text search nor the search form, they managed to visit most other key components, exhibiting behavior that is typically associated with bots (Liu, 2008). This assumption was confirmed by examining the sessions' user agent strings, which connected all 64 sessions with a Java-based browsing component that is typically used for web spiders.

**Table. 1** Site usage by cluster ([1] $\chi^2$ significant $p<0.01$, [2] $\chi^2$ significant $p<0.05$)

| General properties | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
|---|---|---|---|---|
| Number of sessions | 7989 | 2839 | 81143 | 64 |
| % of all sessions | 8.7% | 3.1% | 88.2% | 0.1% |
| Entered through main page[1] | 41.9% | 44.9% | 34.6% | 100% |
| Session duration (min) | 12.6 | 12.2 | 2.5 | 9.1 |
| **% of sessions that changed the interface language to…** | | | | |
| Swedish[1] | 9.7% | 12.5% | 5.3% | 100% |
| English[1] | 8.5% | 8.4% | 7.1% | 100% |
| German[1] | 2.0% | 2.6% | 3.1% | 100% |
| **% of sessions that browsed…** | | | | |
| Accommodation[1] | 80.3% | 24.7% | 26.5% | 100% |
| To Do[1] | 25.8% | 33.2% | 28.3% | 100% |
| To See[1] | 11.7% | 20.7% | 9.2% | 100% |
| Dining[1] | 10.2% | 100% | 8.0% | 100% |
| Service[1] | 6.9% | 14.2% | 8.5% | 100% |
| Communication[1] | 6.0% | 7.7% | 5.8% | 100% |
| Program[1] | 29.2% | 17.7% | 7.2% | 100% |
| **Number of activations (for sessions that browsed a given category)** | | | | |
| Accommodation | 6.4 | 1.0 | 0.5 | 14.0 |
| To Do | 0.8 | 1.5 | 0.7 | 40.0 |
| To See | 0.3 | 0.8 | 0.2 | 51.0 |
| Dining | 0.2 | 8.4 | 0.2 | 40.7 |
| Service | 0.2 | 0.8 | 0.2 | 18.0 |
| Communication | 0.1 | 0.3 | 0.1 | 10.0 |
| Program | 1.8 | 0.8 | 0.1 | 13.0 |
| **Other key actions - % of sessions…** | | | | |
| Viewing a PDF[1] | 2.1% | 2.8% | 6.8% | 0% |
| Viewing a brochure[1] | 5.4% | 7.6% | 3.4% | 100% |
| Ordering a brochure[2] | 0.8% | 2.1% | 0.6% | 100% |
| Return to main page[1] | 18.8% | 18.4% | 14.5% | 100% |
| Visiting "Congress"[1] | 5.3% | 7.2% | 2.4% | 100% |
| Visiting "Tourism"[1] | 11.2% | 10.0% | 6.7% | 100% |
| Visiting "Åre Summer"[1] | 4.6% | 5.1% | 4.2% | 100% |

**Table. 2** Google referrals among clusters ([1] $\chi^2$ significant at $p<0.01$)

| Google referrals | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
|---|---|---|---|---|
| % searched with Google[1] | 41.4% | 44.5% | 47.3% | 0% |
| % referred from Google[1] | 34.0% | 37.8% | 37.2% | 0% |
| % referred to home page[1] | 21.7% | 31.5% | 22.4% | NA |
| Query length (words) | 2.9 | 2.7 | 3.0 | NA |
| **% of Google referrals containing search words that could be associated with…** | | | | |
| Skiing[1] | 1.0% | 0.6% | 3.1% | NA |
| Activities[1] | 3.0% | 2.1% | 6.6% | NA |
| Accommodation[1] | 10.5% | 1.3% | 6.0% | NA |
| Dining[1] | 0.9% | 13.5% | 2.1% | NA |
| Shopping[1] | 0.2% | 0.5% | 1.4% | NA |
| Services[1] | 17.5% | 12.3% | 14.0% | NA |
| Attraction[1] | 0.9% | 0.7% | 1.6% | NA |
| Non-Specific[1] | 91.2% | 95.2% | 83.9% | NA |

Importantly, as shown in Table 2, the analysis of Google referrals to the gateway demonstrates that there is a strong correlation between the categories assigned to search words and the categories visited during the session itself. Both cluster 1 and 2 contained many sessions that were referred on the basis of non-specific search terms

and were, thus, more likely to be referred to the main page. It is surprising that sessions in cluster 3 were so short although many of the search terms could be associated with categories corresponding to the information offerings of the website. As expected, no session in cluster 4 was referred to the site by Google.

When examining the use of search functionality offered by the site itself (Table 3), sessions in cluster 3 were found to be significantly more likely to use the internal text search while they were less likely to use the parametric search form.

**Tab. 3** Internal search and other functionality among clusters ([1] $\chi^2$ significant $p<0.01$)

| **Internal text search** | **Cluster 1** | **Cluster 2** | **Cluster 3** | **Cluster 4** |
|---|---|---|---|---|
| % sessions that used it[1] | 9.4% | 9.5% | 15.0% | 0% |
| First used after (min) | 3.4 | 3.7 | 1.3 | NA |
| Query length (words) | 1.7 | 1.6 | 1.5 | NA |
| **% of internal text searches that could be associated with…** | | | | |
| Skiing | 5.5% | 8.1% | 7.9% | NA |
| Activities | 5.1% | 4.4% | 4.8% | NA |
| Accommodation[1] | 22.1% | 17.7% | 15.7% | NA |
| Eating[1] | 4.0% | 11.1% | 2.4% | NA |
| Shopping[1] | 0.1% | 2.6% | 1.5% | NA |
| Services[1] | 4.9% | 7.7% | 10.2% | NA |
| Attraction[1] | 2.9% | 2.6% | 5.1% | NA |
| Non-Specific | 30.2% | 25.5% | 27.8% | NA |
| **Parametric search form** | | | | |
| % sessions that used it[1] | 21.0% | 18.8% | 6.0% | 0% |
| First used after (min) | 5.0 | 4.3 | 1.5 | NA |
| Average number of uses | 4.4 | 4.9 | 1.9 | NA |

In both instances, sessions in cluster 3 that used search functionality were more likely to do so earlier in the session (see the *first used after* variables). Interpretations of this difference could include that these sessions had difficulty locating relevant information using category-based site structure or that the information offerings of the parametric search facility (that allows users to specify specific search criteria like type of accommodation or activities) did not match their actual information needs. Given that once again most of cluster 3's internal text search terms could be associated with a category, it is reasonable to assume that this is an indication that the site would benefit from additional content, particularly in the areas of *skiing*, *accommodation* and *service*.

We complete our analysis by returning to the search terms used for both Google referrals and internal text searches. Figure 4 lists the 50 most common key terms used in Google searches for each cluster, with size indicating a term's relative frequency. Non-differentiating terms were excluded since such terms (e.g. *åre*) are a de facto requirement for referrals to the site (Wolk & Wöber, 2009). Interestingly, each cluster contains a number of topical terms specific to that cluster. For example, cluster 2, clearly interested in dining, contains terms such as *pub*, *konditori* and *restaurang*.

Turning our attention to the terms used in internal text searches (Figure 5), it is obvious that the terms refer to more specific needs and are in general more diverse.

a. Cluster 1

accommodation accomodation afterski areturistbyra åreturistbyrå att bo boende bröllop camping com fiske fjällby fjällgård göra händer holiday höstmarknad hotel hotell hyra in inn jämtland karta lägenhet lägenheter nattliv och övernattning pensionat personer privatstugor rum ski skidstuga storulvån stuga stugby stugor timmervillan turistbyra turistbyrå turistbyrån turistinformation uthyres uthyrning vad vandrarhem vecka

b. Cluster 2

affärer armfeldts äta att bar bistro black cafe café club dahlboms dds fjällby göra helpension hotell ica janne julbord kabinbana kabinbanan karta knuten konditori krog mat meny morsel oliven på pub raw restaurang restauranger runt schaffer sheep stormköket supper tännforsen taxi timmerstugan toppstugan tts turer turistbyrå turistbyrån turistinfo våfflor xc3

c. Cluster 3

anjans att bibliotek bo boende buss camping com fiske fjällby fjällstation från göra holiday höstmarknad hotel hotell hyra ica inn islandshästar jämtland julfirande kabinbanakabinbanan karta kyrka längdåkning längdspår och öppettider på restaurang restauranger shop skalstugan skiduthyrning skoter skoterleder skoteruthyrning stuga stugor till transfer turistbyrå turistbyrån turistinformation vandra vandrarhem vandring

**Fig. 4** Dominant Google search terms by cluster

a. Cluster 1

afterski alpina åregårdarna åregården åresjön äta bastu bergbanan bjørnen bo boende bygget club diplomat fiskecamp fjällby fjällby fjällbyn fjällbys fjällgård fjällgården fjällhotell fjällstuga fjällvärlden gästhuset holiday hotel hotell internet jope kabinbanan karta kommunikationer kvm kyrka liftkort mörviksgården oviksfjällen pistkarta radio renen röding rum rustika service ski skidliftar snasahögarna tännforsen toaletter

b. Cluster 2

after anaris åregården åresjön armfeldts äta bädds bar bastu bastuanläggning bergbanan blåhammarens bo bustamoen butiker bykrog bykrogen club continental edhsgården fäviken fiskefiskecamp fj fjällbys fjällgården fjällhotell grappa hummelstugan inn julbord kabinbanan karta kläppen knuten krog liftkort lillåstugan rautjoxa renfjället restaurang restauranger service shopping skoteruthyrning sträcker timmerstugan tottebo våfflor worsens

c. Cluster 3

afterski alpina åregården åresjön äta bastu bergbanan bo bubbelpooler churchill club dammån fj fjällbyn fjällbys fjällgård fjällhotell fjällstationer fjällstugor hälsocentral hälsocentralen heliski hotell hundspann kabinbanan karta klockstapeln kommunikationer kvissleströmmarna kvm lägenhet liftkort monte oviksfjällen röding service sevärdheter skalstugan ski skidliftar skoter skoteruthyrning slalom storulvåns tännforsen toaletter totto väder villa winston

**Fig. 5.** Dominant internal search terms by cluster

The dominating types of search words used by cluster 1 in internal text searches are associated with accommodation. Search terms such as *fjällbyn* and *åregården* are names of hotels, while *bo* and *hotel/hotell* refer to accommodation category more generally. Similarly, in cluster 2 the dominating search terms refer to the dining category, either by using the names of restaurants (e.g. *worsens, fjällgården)* or by using more generic search words (e.g. *restauranger, äta, julbord)* associated with this information category. Summarizing, the different clusters clearly represent user groups with distinct information needs. However, the fact that the vast majority of users (i.e. cluster 3) remained on the site for such a short period of time, particularly in comparison to the presumably more successful clusters (i.e. clusters 1 and 2), is an indication that the site may need to be reorganised and expanded to fit different customer groups and information needs. Typically, in an e-marketing context the customer initiates a search for information on products and services for which they have already formed an interest (Biswas & Krishnan, 2004). Consequently, it is important to satisfy customer needs once they are using a web site. Failure to do this may reflect badly on the destination brand or signify lost business opportunities.

## 5 Conclusions

This paper introduced a workflow for utilizing standard web server log data in WUM. In order to show how such a workflow might be applied to a real-world website, we considered the example of the *visitare.se* tourism destination gateway and demonstrated how usage data can be extracted and processed. In particular we focused on examining search terms as well as identifying differences between homogenous user groups revealed by unsupervised learning techniques. When examining the results of our investigation, it is evident that different user groups approach the site with significantly differing information needs (Bhatnagar & Ghose, 2004).

More precisely, our results indicate that the majority of users (i.e. cluster 3) spend rather little time on the site, perhaps signifying that their specific information needs have not been met. Admittedly, at such an early stage in our research it is not possible to ascertain the causes for the differences between the clusters, something that remains for future work. Nevertheless, knowledge about specific user groups is crucial for understanding the relationships between (potential) tourism products, users and related information categories. Finally, the diversity of search words and the value of Google as a source of referrals strongly underline the importance of proper search engine management.

## References

Biswas, A. & Krishan R. (2004). The Internet's impact on marketing, *Journal of Business Research*, 57 (7): 681-684.

Bhatnagar, A. & Ghose, S. (2004). Segmenting consumers based on the benefits and risks of Internet shopping, *Journal of Marketing Research*, 40 (2): 235-243.

Cho, V. & Leung, P. (2002). Towards using knowledge discovery techniques in database marketing for the tourism industry, *Journal of Quality Assurance in Hospitality & Tourism*, 3(3): 109-131.

Dias, J. G. & Vermunt J. K. (2007). Latent class modeling of website users' search patterns: Implications for online market segmentation, *Journal of Retailing and Consumer Services*, (14): 359-368.

Hastie, T., Tibshirani, R. & Friedman, J. (2009). The elements of statistical learning – Data mining, inference and prediction (2nd ed.), New York, Springer.

Honda, T., Yamamoto, M. & Ohuchi, A. (2006). Automatic Classification of Websites based on Keyword Extraction of Nouns, In, Hitz, M., Sigala, M., Murphy, J. Eds.), *Information and Communication Technologies in Tourism 2006* NY: Springer: 263-272

Jolliffe, I. T., (2002). Principal Component Analysis. Springer-Verlag. USA.

Larose, D.T. (2005). *Discovering knowledge in data – An introduction to data mining*. John Wiley & Sons, New Jersey.

Liu, B. (2008). *Web Data Mining - Exploring hyperlinks, contents and usage data* (2nd ed.), Springer, New York.

Mobasher, B. (2008). Web Usage Mining, In: Liu, B. (ed.) *Web Data Mining- Exploring hyperlinks, contents and usage data* (2nd ed.) Springer, New York, pp 449-483.

Murphy, J., Hofacker, C.F., & Bennett, M. (2001). Website-generated Market-Research data: Tracing the tracks left behind by visitors, *Cornell Hotel and Restaurant Administration Quarterly*, 42(1): 82-91.

Olmeda, I. & Sheldon, P.J. (2002). Data Mining Techniques and Applications for Tourism Internet Marketing, *Journal of Travel & Tourism Marketing*, 11(2/3): 1-20.

Orlando, S. & Silvestri, F. (2009) Query Log Analysis for Enhancing Web Search, IEEE/WIC/ACM International Conference on Web Intelligence, Milano, Italy.

Pelleg, D., Moore, A.W. (2000). X-means: Extending K-means with Efficient Estimation of the Number of clusters, *Proceedings of the Seventeenth International Conference on Machine Learning*. pp727 – 734. USA.

Pyle, D. (1999). *Data preparation for data mining*. New York, Morgan Kaufmann Publishers.

Scharl, A., Wöber, K. & Bauer, Ch. (2004). An integrated approach to measure web site effectiveness in the European hotel industry, *Journal of Information Technology and Tourism,* 6(4): 257-271

Schegg, R., Steiner, Th., Gherissi-Labben, T. & Murphy, J. (2005). Using Log-File Analysis and Website Assessment to Improve Hospitality Websites, Frew, A.. (Ed.) *Information and Communication Technologies in Tourism 2005*. New York, Springer, 566-576.

Tichler, G., Grossman, W. & Werthner, H. (1999). Using Data Mining in Analysing Local Tourism Patterns. In Buhalis, D. & Schertler, W. Eds., *Information and Communication Technologies in Tourism 1999* Vienna: Springer: 1-11.

Werthner, H. & Ricci, F. (2004). E-commerce and tourism. *Communications of the ACM*, 47(12): 101-105.

Wolk, A. & Wöber, K. (2009). A Comprehensive Study of Info Needs of City Travellers in Europe, *Journal of Information Technology and Tourism*, 10(2): 119-131.

## Acknowledgement

Parts of this work have been financed by the ÖNB grant no. 13.000 of the Austrian National Bank Jubilee Foundation and by the EU Structural Fund objective 2 project no. 39736, Sweden.