APPLICATION OF WEB MINING IN CREATING A LONG

LASTING CUSTOMER RELATIONSHIP

Celina Olszak, Śląska Wyższa Szkoła Zarządzania im. gen. J. Ziętka w Katowicach
(The General Jerzy Ziętek Silesian School of Management),
ul. Krasińskiego 2 40-952 Katowice
olszak@swsz.katowice.pl

Kamila Bartuś, Śląska Wyższa Szkoła Zarządzania im. gen. J. Ziętka w Katowicach,
(The General Jerzy Ziętek Silesian School of Management),
ul. Krasińskiego 2 40-952 Katowice
kdymek@swsz.katowice.pl

Tomasz Bartuś, Śląska Wyższa Szkoła Zarządzania im. gen. J. Ziętka w Katowicach,
(The General Jerzy Ziętek Silesian School of Management),
ul. Krasińskiego 2 40-952 Katowice
tbartus@swsz.katowice.pl

ABSTRACT

The article attempts to characterise the most important Web Mining techniques. Benefits that result from implementing the techniques in question to improve customer relations are presented. Subsequently, a brief overview of tools – offered by selected companies - that are used to extract data from the World Wide Web is given.

KEY WORDS:

Web Mining, Web Content Mining, Web Structure Mining, Web Usage Mining, Text Mining, CRM.

1. INTRODUCTION

Customer Relations Management (CRM) is a business strategy that aims at sustaining current contacts with customers and creating new relations between a company and its customers. Such relations may take a form of long-lasting partnership when a company is able to collect, process and analyse the largest possible amount of information on its customers appropriately [18].

Different analyses of products, services and customers may – more and more frequently – be carried out on the basis of data obtained by means of the Internet since Internet websites make up a perfect environment for determining preferences of customers' behaviours. They enable to do research, monitor research in the real time and correlate research with the sales results. Analyses of the way a customer moves around websites together with websites thematic scope, sequence and a number of hits may provide a lot of valuable information. The World Wide Web is an incredibly rich repository of diversely structured data. Automatic extraction of information from this kind of data requires implementation of appropriate methods and techniques. Web Mining techniques may serve well here as they allow for gaining and analysing much information from WWW including information on customers as well.

2. NATURE AND FORMS OF WEB MINING TECHNIQUES

Web Mining is a technique that aims at automatic extraction of information from WWW sites and WWW documents [3][12]. Major tasks of Web Mining should include:

- extracting and processing of data that comes from WWW and e-mail text documents;
- sorting selected information extracted from WWW resources;
- discovering major patterns on single WWW sites and all websites by means of the following techniques: classification, regression, finding sequence patterns, grouping and finding exceptions and deviations; and
- analysing patterns obtained [12][4].

Taking the way and scope of searching for data in the Web into consideration, Web mining is divided into Web Content Mining, Web Structure Mining and Web Usage Mining. Web Content Mining techniques are responsible for automatic searching and reporting of WWW sites and documents and for analysing Internet databases. Web Structure Mining analyses a structure of the hyperlinks within the Web itself. On the other hand, Web Usage Mining allows for recognising patterns of users' interacting with the Web and their logging to WWW sites (Figure 1).
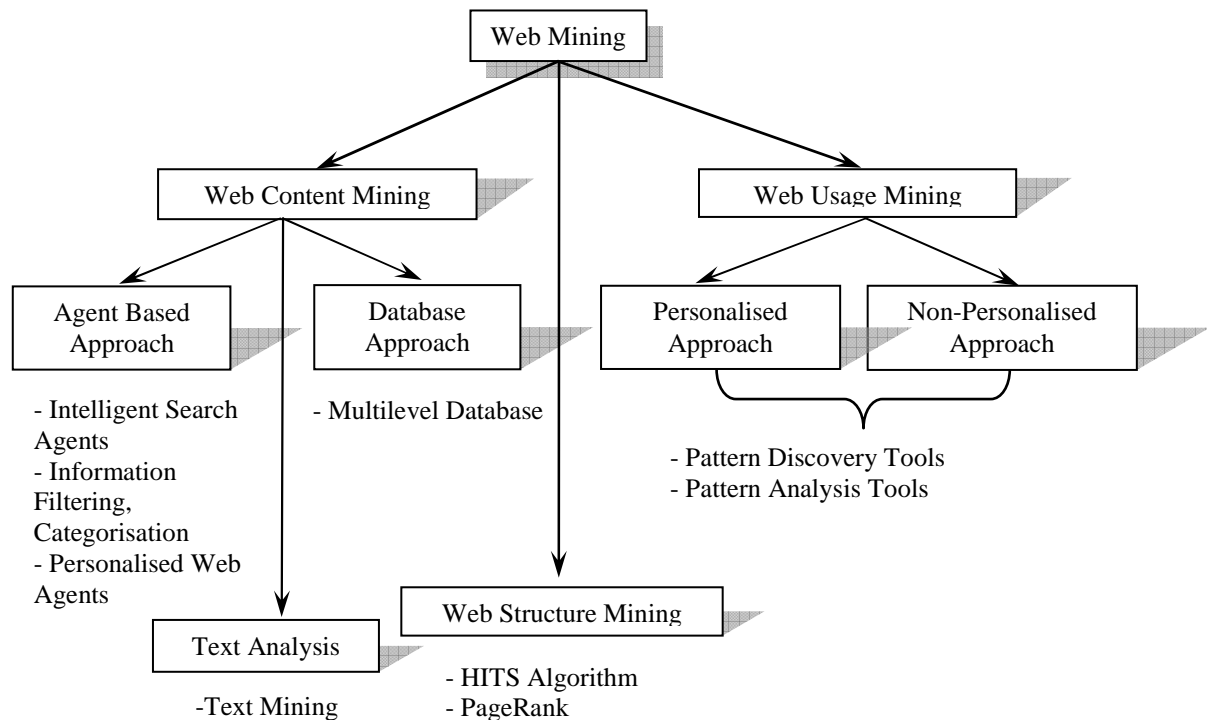
```
                          ┌─────────────┐
                          │ Web Mining  │
                          └─────────────┘
              ┌──────────────────┴──────────────────┐
    ┌──────────────────┐                    ┌──────────────────┐
    │ Web Content      │                    │ Web Usage Mining │
    │ Mining           │                    │                  │
    └──────────────────┘                    └──────────────────┘
      ┌───────┴───────┐                       ┌────────┴────────┐
┌──────────┐  ┌──────────┐            ┌──────────────┐ ┌──────────────┐
│ Agent    │  │ Database │            │ Personalised │ │ Non-         │
│ Based    │  │ Approach │            │ Approach     │ │ Personalised │
│ Approach │  │          │            │              │ │ Approach     │
└──────────┘  └──────────┘            └──────────────┘ └──────────────┘
```

- Intelligent Search Agents
- Information Filtering, Categorisation
- Personalised Web Agents

- Multilevel Database

- Pattern Discovery Tools
- Pattern Analysis Tools

```
                     ┌──────────────────────┐
                     │ Web Structure Mining │
                     └──────────────────────┘
        ┌─────────────────┐
        │ Text Analysis   │
        └─────────────────┘
```

- HITS Algorithm
- PageRank

-Text Mining

Figure1 . Web Mining Structure
Source: Own on the basis of [6][19]

2.1 Web Content Mining

Web Content Mining aims at extracting information from the WWW resources to process it into useful knowledge. Data structure in the WWW environment is heterogeneous. As a result, there are enormous difficulties in searching and processing of data. Traditional WWW search engines including Google, Altavista or Yahoo successfully search for data on a particular subject. However, they do not provide information on the structure of searched documents, and do not apply advanced mechanisms of documents categorising, filtering and interpreting. The above problem may be partly solved by means of Web Content Mining techniques that borrow some solutions from the agent based, database and text analysis approaches.

In the Agent-Based Approach with relation to Web Content Mining particularly much attention is paid to Intelligent Search Agents, Information Filtering and Categorisation together with Personalised Web Agents [6].

Intelligent Agents are responsible for searching relative information by referring to knowledge of a particular domain and users' profiles. A pre-prepared data base is used to achieve that. In case unknown knowledge is encountered, agents may learn and analyse a new domain by means of interactions with users [6]. An example of such a learning system may be provided by IntellShoper [17].

In Web Content Mining, hierarchical clustering techniques are used to search for information. The techniques in question allow for automatic document filtering and categorisation. Their hierarchy is set on the basis of the link structures and document content. Such solutions have been applied in, inter alia, HyPursuit and Bookmark Organizer [25] [16].

Personalised Web Agents prove particularly useful in discovering information. These techniques are user-oriented and they enable to learn users' behaviours, and preferences in searching for information or doing shopping, etc. [6]. An example is given by, for instance WebWatcher [2]. Intelligent Agent techniques may also be helpful while updating document databases. By means of the agents in question in certain periods of time WWW sites are verified to check whether they are updated [21].

Techniques of discovering information that root from the Database Approach may be utilised to integrate and organise heterogeneous information coming from numerous sources in order to obtain a compact and multilayer structure of resources. For example, the lowest level of such a structure may be used to store semi-structured information, e.g. in form of HTML documents. Subsequent layers may be used to store metadata of the above documents that are organised in relation and object databases. The last level should be reserved for new metadata [6].

Text documents (stored in the HTML standard) form a large part of the World Wide Web resources. In this case, to discover new information and knowledge, it is necessary to use different algorithms of the text analysis. Discovering interdependencies that are observed in text documents is possible by employing, inter alia, text mining techniques. Text mining techniques use association of key words or similarity to document patterns. Text mining techniques are also used to determine relations between documents in a particular set, combining free texts, etc. [15] [11] [19].

2.2 Web Structure Mining

Web Structure Mining techniques are used to discover a model of a structure of interdependencies that are observed in WWW. The model in question bases on the hyperlink typology and hyperlink description. It may be used to categorise websites and generate information on similarities and relations that are observed between different websites [12]. Some algorithms including HITS (Hyperlink Induced Topic Search) and PageRank make it possible to build WWW typology model that is used to evaluate quality and analyse page positioning in the website.

HITS algorithm developed by Jan Kleinberg provides for an assumption that websites are interconnected to resemble a graph. In the graph, websites are nodes of the graph and hyperlinks are edges of the graph. The algorithm examines mutual interconnections between websites [7]. If one website has a hyperlink to the other, it is possible to assume with high probability that the website is well constructed. As a result a user who searches for a particular issue, service or goods is given a set of appropriately selected websites.

PageRank algorithm (created by Larrg Page and Sergey Brin - founders of Google) involves random link visiting and counting visited websites. More hits of a particular website indicate its higher popularity and – simultaneously – higher position in the ranking. PageRank level of a particular website does not only depend on the number of links to the very website but also on the PageRank value of websites where particular links are to be found [26]. For instance, www.google.pl enjoys PageRank rate that equals 8. However, in case of www.eactive.pl PageRank rate is 5.

2.2 Web Usage Mining

Web Usage Mining techniques are used for automatic discovering WWW users' behavioural patterns[6]. There are two approaches in the Web Usage Mining: personalised and non-personalised. In the former particular user's behaviour is analysed and the user's profile is obtained. On the other hand, in the latter more extensive group of users is subject to observation. All analyses in the Web Usage Mining are carried out on the basis of data that mainly comes from log files, sessions, cookies and users' profiles. Users' activeness, moves and time of visits on the WWW is recorded [12]. Such information may be used to determine customer value, create effective marketing campaigns, etc. An exemplary solution may be provided here by NetTracher developed by Sane Solution. System. NetTracher is equipped with – first of all – a module to discover and acquire data from servers logs and it allows for counting the most and least popular WWW sites, number of hits and so on [28].

Additional analyses, interpretations and visualisations of certain patterns of WWW users may be obtained due to OLAP technologies. WebHound by SAS offers a possibility to store historical data,

generate structures of OLAP multidimensional cubes and visualise and present obtained results in form of reports [10] [29].

3. WEB MINING TECHNIQUES USED IN CUSTOMER RELATIONS MANAGEMENT

Customer relations management, as it has already been mentioned, is an organised attempt to improve identification, acquisition and maintenance of a customer. CRM aims at improvement in the level of customer satisfaction, loyalty and profits that originate from the mutual co-operation [24]. The concept predominantly serves the following purposes [13]:
- providing all company employees with exact information so they could answer all customers' questions;
- obtaining complete knowledge on a customer, e.g. what they have purchased, what offers they have taken advantage of; and
- overall service of contacts with customers including, inter alia, telephone, fax, e-mail and the Internet.

Practice shows that effective relations with customers should be created on the basis of up-to-date, adequate and reliable information. To create a complete model of a customer it is also necessary to possess some knowledge on forecasting and analysing customers' needs. Only then it is going to be possible to offer customers suitable packages of products and services.

Nowadays, more and more customers use Internet portals and shops. They leave a lot of valuable information there. The information may then be utilised by companies wishing to sell their products. The way a customer moves around the website together with products added to a shopping cart are only exemplary pieces of information that are stored in log files of WWW servers.

Analyses of information to be found in WWW server logs belong to one of Web Mining techniques applications. The techniques use methods of discovering frequent navigation paths, association principles, behavioural patterns, classification and grouping [12].

The former of the methods mentioned above is used to analyse navigation of users in environments where there are many documents that are interconnected by a network of mutual links [5] [27]. Such links are the easiest to present by means of a graph where WWW sites are nodes of the graph and hypertext links are edges of the graph. Such a structure allows for analysing all users' transactions or profiles and discovering paths that lead to purchasing of a given product. If - instead of links - numbers of users who transferred from one site to another are placed in nodes, then it is possible to define the most frequently visited paths that lead to particular sites. Such activities do help reach the highest number of customers, e.g. by placing advertising banners promoting particular products on the sites in question [9] [8].

Principles of association may support planning marketing strategies on WWW sites. Discovering such principles in server logs is carried out to obtain information on sets of sites users refer to. An example may be provided by a principle according to which 60% of users who have visited www.abc.pl simultaneously have visited www.xyz.pl [27]. Filtering is a method that is based on association principles. Amazon.com is the most famous company that uses this method. On the basis of purchases and a geographical location of a particular customer the method forecasts what other readers are going to buy (those who live nearby and have similar interests) [9].

While planning a marketing strategy it is also possible to apply sequence patterns [1][22]. Application of such methods may help determine characteristics of all customers who have visited a particular website or a sequence of websites in a given time. Such research helps forecast network overload in the particular time, adjust site content to a group of portal visitors, carry out optimisation of time sequence of the whole website system, and this way optimise a strategy of a contact with customers [8].

Grouping aims at dividing a set of objects into groups so that similarity between objects in the same group could be as high as possible. In case of objects representing different groups, such similarity is supposed to be as low as possible. Grouping is used to distinguish groups of customers who enjoy the same interests. It serves e.g. developing a more successful offer [27] [14].

As a result of classifying a user to a particular group of customers (who represent similar interests) it is possible to personalise dynamic advertising banners and promotions. For classification purposes neuronal networks and decision trees are used. The method mainly serves supporting campaigns that aim at transforming users who 'view' into those who 'buy', and those who 'buy' into those who 'come back'.

The methods of data exploration described above allow for better designing Internet websites, successful addressing advertisements to customers and building marketing strategies. Such activities undoubtedly lead to improvement in relations with customers.

4. CASE STUDIES OF UTILISING WEB MINING TO IMPROVE CUSTOMER RELATIONS

At present, numerous companies offer comprehensive solutions that utilise Web Mining techniques in order to improve relations with their customers. Providers of such solutions include, inter alia, WebMiner and SPSS.

WebMiner's tools have remarkably influenced marketing strategy development of a company that trades in air-conditioners. The main source of information predominantly included WWW data on customers and a CRM database (storing information on 25,221 customers). The data was combined with aggregated demographic data in three intersections:
- an analysis of post codes;
- an analysis of lifestyles; and
- an analysis of properties.

Having applied a decision tree three groups of customers were identified. Those customers were characterised by three times higher than the average willingness to purchase air-conditioners. The groups in question included immigrants who inhabited the eastern coast - a group of well educated and highly paid, young Spanish speaking families of immigrants. Besides, it turned out that the customers in question mainly lived in some umpteen centres of New York and New Jersey States where housing facilities predominantly originated from before fifties and were deprived of central air-conditioning systems. In majority of cases target customers had no cars. As a result of works undertaken WebMiner recommended the interested company to give up radio commercials and switch to bus and underground advertising. Banner advertising was suggested in case of websites oriented at young inhabitants of the eastern coast. Moreover, a Spanish language version of the Internet shopping facility was to be prepared [23] [30].

Application of tools by SPSS shows how it is possible to adjust a product to individual needs of a customer on the basis of data that is accumulated in the process of registering on WWW sites. By means of this tool a Japanese company Sofmap delivering computer hardware succeeded in creating a model that was used to personalise contents of the Softmap website. As a result of analyses carried out, it was possible to provide each customer who visited their website with a customised offer that would likely be accepted by the very customer. Due to application of Web Mining techniques the company tripled its profits originating from the Internet and increased the number of its website visits from eighteen million of hits monthly to over thirty five million [31].

Jubii – the largest Danish Internet portal – was looking for the way to increase frequency of visits to their websites and raise users' loyalty. By means of Clementive tools provided by SPSS they undertook realisation of the project that was supposed to profile users. Taking advantage of logs four profiles for each users were created. The profiles encompassed categories of websites visited during a week and at weekends, and time of visits (working and free days). Updated and detailed (at every visit) profiles of Internet users were used to create a base for a prediction model that subsequently helped create an optimisation programme for advertising services. As a result of the actions described above there was some rise in number and frequency of visits at Jubii sites. Moreover, advertising and sponsoring increased [31].

5. CONCLUSIONS

Creation of individual and long-lasting relations with customers have become a priority for companies. Practice shows that a significant number of companies use the Internet to get familiarised with customers' preferences, find interdependencies in customers' behaviours and analyse customers. Better understanding of customers' preferences and habits may be supported by Web Mining techniques that involve discovering association principles and sequence patterns, classifications and grouping, and finally discovering of navigation paths. Appropriate utilisation of the above methods is a key to build personalised WWW sites and offers, and it allows for running successful advertising campaigns.

5

REFERENCES

[1] Agrowal R., Srikan R.: *Mining Sequential Patterns,* Conference on Data Engineering (ICDE), Taiwan, March 1995

[2] Amstrong R., Freitag D., Joachims T., Mitchell T.: *Webwatcher: A learning apprentice for the World Wide Web,* In Working Notes of the AAAJ Spring Symposium: Information Gathering from Heterogeneous, Distributed Envirnoments, Stanford University, 1995

[3] Berendt B., Hotho A., Stumme G.: *Towards Semantic Web Mining,* Springer-Verlag Berlin Heidelberg, 2002

[4] Chakrabarti S.: *Mining the Web, Discovering Knowledge from Hypertext Data,* Morgan Kaufmann Publishers, Elsevier Science, 2003

[5] Chen M. S., Park J. S., Yu P. S.: *Effectin Data Mining for Path Traversal Patterns,* IEEE Transactions on Knowledge and Data Engineering, March 1998

[6] Colley R., Mobsher B., Srivastava I.: *Web Mining: Information and Pattern Discovery on the World Wide Web,* IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), 1997

[7] Czyżowicz M.: *Badania porównawcze metod inteligentnej nawigacji i adaptacji algorytmów optymalizacyjnych do zadań nawigacji w zbiorach dokumentów tekstowych,* Praca magisterska pod kier. prof. nzw. PW dr hab.. Inż.. Kłopotka M. A., Politechnika Warszawska, Warszawa 2003

[8] Dudzicz M.: *Zastosowanie metod eksploracji danych dla handlu elektronicznego,* [w:] Komputerowo Zintegrowane Zarządzanie pod red. Ryszarda Knosali, WNT, Warszawa 2004

[9] Dyche J.: *CRM relacje z klientami,* Helion, Gliwice 2002

[10] Dyreson C.: *Using on incoplete data cube as a summary data sieve,* Bulletion of the IEEE Technical Committee on Data Engineering, March 1997

[11] Filipowski A.: *Jak zaoszczędzić na czytaniu?,* Gazeta IT 3(22), Poznań, luty 2004

[12] Kosala R., Blochecl H.: *Web Mining Research: A Survery,* SIGKDD Explorations,Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining 2(1), 2000

[13] Kotowski M.: *Oswoić klienta,* PCKurer nr 26, 1996

[14] Kulpa P.: *Visitors Relationship Management. Nowe wyzwanie dla hurtowni danych,* PCKurier 24/2201

[15] Lula P.: *Text Mining jako narzędzie pozyskiwania informacji z dokumentów tekstowych,* [w:] Data Mining: poznaj siebie i swoich klientów, Statsoft, Warszawa-Kraków, 2005

[16] Maarch Y. S., Ben Shavl I. Z.: *Automatically organizing bookmarks per content,* In Proc. Of 5th International World Wide Web Conference, 1996

[17] Menczer F., Monge A., Street W.N.: *Adoptive Assistants for Customized E-shopping,* IEEE Intelligent Systems, 2002

[18] Olszak C., Bartuś K., Bartuś T.: Wykorzystanie wybranych technik drążenia danych w systemach zarządzania z klientami, [w:] Komputerowo Zintegrowane Zarządzanie, pod red.: R. Knosali, Oficyna Wydawnicza Polskiego Towarzystwa Zarządzania Produkcją, Opole, 2006

[19] Rajman M.: *Text Mining: Natural Language techniques and Text Mining applications,* In Data mining and reverse engineering: searching for semantics, SPACCAPIETRA S. and MARYANSKI F., Eds. Chapmann and Hall, London, 1997

[20] Scime A.: *Web Mining: Applications and Techniques,* Published in the United States of America by Idea Group Inc, 2005

[21] Solberg E. W.: *Towards Comprehensive Web Search,* Doctoral Dissertation University of Washington, 1999

[22] Sriknat R., Agrowal R.: *Mining Seqvential Patterns: Generalisations and Performance Improvements,* Conference on Extending Database Technology (EDBT), France, March 1996

[23] Stawowy A.: *Data Mining w systemach CRM,* www.zarz.agh.edu.pl/astawowy.Artykul/dm_crm.pdf

[24] Tiwana A.: *Przewodnik pozarządzaniu wiedzą. E-biznes a zastosowanie CRM,* Placet, Warszawa, 2003

[25] Weiss R., Velez B., Sheldon M.A., Namprempre C., Szilagyi P., Duda A., Gifford D. K.: *Hypursiut: a hierarchical network search engine that exploits content-link hypertext clustering,* In Hypertext'96: The Seventh ACM Conference on Hypertext, 1996

[26] Węcel K.: *Jek Google to robi?,* Gazeta IT nr 9, 2005

[27] Wojciechowski M.: *Odkrywanie wzorców zachowań użytkowników WWW,* [w:] Materiały konf. POLMAN'99, OWN, Poznań 1999

[28] www.sane.com

[29] www.sas.com

[30] www.webminer.com

[31] www.spss.com