STUDIA INFORMATICA 2002
Volume 23 Number 2B (49)

Anna ZATWARNICKA
Politechnika Opolska, Wydział Elektrotechniki i Automatyki

WYKORZYSTANIE DATA MINING DO POPRAWY UŻYTECZNOŚCI PRZEGLĄDAREK INTERNETOWYCH I STRUKTURY STRON WWW

Streszczenie. Obecnie w sieci WWW występuje tak wielka ilość informacji, że coraz bardziej potrzebne staje się oprogramowanie, które pomogłoby nad nią zapanować. Serwisy webowe również potrzebują oprogramowania wspomagającego odpowiednie jej prezentowanie. Dokument ten ma przybliżyć pewne rozwiązania, dotyczące tworzenia profili użytkownika oraz poprawiające funkcjonalność stron WWW.

APPLY OF DATA MINING TO IMPROVE EFFICIENCY OF WEB BROWSERS AND STRUCTURES OF WEB PAGES

Summary. Nowadays is difficult to find right piece of information on World Wide Web. There is need to build application which can help user to obtain significant and proper information. Web services also need software, which effective assist presentation of information. This paper approaches some solutions, which were created in order to make user's profile and solutions, which help functionality of web browsing.

1. Wprowadzenie

Internet stał się dzisiaj swoistą miarą postępu technicznego. Wzrasta liczba serwisów internetowych: edukacyjnych, tematycznych, komercyjnych. Stosunkowo łatwo jest rozmaite informacje w sieci umieścić – jeszcze łatwiej je przeczytać. I do tego właśnie Internet używany jest najczęściej – do wymiany informacji. Inne usługi, takie jak transfer plików (ftp), zdalne logowanie się (telnet) odchodzą w zapomnienie i z rzadka tylko są używane

przez bardzo wtajemniczonych. Użytkownicy sieci (internauci) korzystają za pośrednictwem programów nazywanych przeglądarkami internetowymi z sieci WWW po to, by znaleźć interesujące ich informacje zapisane w plikach html (ang. *Hypertext Meta Language*, meta język służący do tworzenia stron internetowych), tekstowych, graficznych i muzycznych. Pliki te mogą znajdować się na różnych komputerach, a powiązane są poprzez system tzw. dowiązań (linków, ang. *links*). Pomocą w ich poszukiwaniach służyć mogą tzw. wyszukiwarki internetowe.

2. Wyszukiwarki internetowe

Wyszukiwarki internetowe są to specjalistyczne systemy, służące do wybierania i umiejscowienia informacji dostępnych w sieci WWW. Są dostępne pod odpowiednimi adresami WWW. W typowej wyszukiwarce internetowej można wyróżnić trzy główne komponenty [6]:

- ludzki lub automatyczny komponent odpowiedzialny za przeszukiwanie sieci i indeksowanie jej zawartości
- bazę danych, gdzie owe indeksy są przechowywane i
- mechanizm wyszukujący, który umożliwia użytkownikowi przeszukiwanie bazy indeksów w poszukiwaniu odpowiednich informacji;

Mechanizm wyszukujący umożliwia użytkownikowi podanie jednego lub kilku słów kluczowych, a czasami nawet całej frazy. Są one następnie dopasowywane przez mechanizm wyszukujący do zawartości indeksów w bazie danych. Każdy indeks ma określoną wagę, pomocną w doborze najbardziej odpowiedniego indeksu. Po znalezieniu kilku najlepszych indeksów mechanizm wyszukujący zazwyczaj zwraca nam tytuł, krótkie podsumowanie i adres internetowy strony odpowiadającej znalezionemu indeksowi (URL, ang. *Uniform Resource Location*). Niektóre systemy wyszukujące umożliwiają użytkownikowi wysyłanie zapytań do wielu mechanizmów wyszukujących. Systemy takie nazywane są Meta-wyszukiwarkami (ang. *Meta Search Engines*).

Znajomość adresów wyszukiwarek internetowych jednak nie wystarcza, nie zawsze bowiem sugerują one linki do odpowiednich stron – o tym wie każdy, kto choć raz chciał znaleźć coś w sieci. Przyczyny niepożądanego działania wyszukiwarek są dwie: po pierwsze, zbyt prosto indeksowana baza danych. Wyszukiwarka przeglądając sieć analizuje słowa kluczowe występujące na stronie. Zazwyczaj nie bada ich znaczenia ani też kontekstu, w którym te słowa występują. Po drugie: nie zawsze poprawne słowa wpisane przez użytkownika. Należałoby się oburzyć: przecież użytkownik, jeśli chce coś znaleźć, to chyba

wie, czego chce? Otóż nie zawsze. Jeśli weźmiemy pod uwagę występujące prawie w każdym języku synonimy (jedno znaczenie może być opisane za pomocą wielu słów) i polisemy (jedno słowo ma kilka znaczeń), przestaniemy dziwić się, że przeglądarka nie będzie w stanie domyśleć się,o co nam chodzi. Problem synonimów można w bardzo łatwy sposób rozwiązać dołączając do przeglądarki tezaurus. W przypadku polisemów można użyć sprzężonej z wyszukiwarką bazy danych, zawierającej ich znaczenia [6]. Użytkownik mógłby wybrać odpowiednie, ale i tak nie miałby pewności, że wyszukiwarka przedstawi mu linki do odpowiednich stron. Bo skąd ona ma wiedzieć, w jakim znaczeniu słowo jest na danej stronie użyte? Najlepiej byłoby, gdyby się jednak domyśliła. Można sprawić, by sprzężona z przeglądarką aplikacja uczyła się zainteresowań użytkownika (wtedy zapamięta, w jakim znaczeniu użytkownik używał danego słowa – polisemu), a gdy się nauczy – można utworzyć dla tego użytkownika specjalny profil, z którego będzie zawsze korzystał używając przeglądarki. Do uczenia się takiej aplikacji wykorzystywane są mechanizmy z dziedziny zwanej Data Mining.

3. Data mining

Data Mining można tłumaczyć jako pozyskiwanie (wydobywanie) wiedzy. Dziedzina ta powstała stosunkowo niedawno i często nazywana jest też odkrywaniem wiedzy zapisanej w bazach danych (ang. *KnowledgeDiscowery in Databases - KDD*). U podstaw Data Mining leżą trzy dziedziny: automatyczne uczenie się maszyn (ang. *machine learning*), statystyka i bazy danych. Główne zadanie tej nowej dziedziny to efektywne pozyskiwanie użytecznej wiedzy z ogromnej ilości danych.

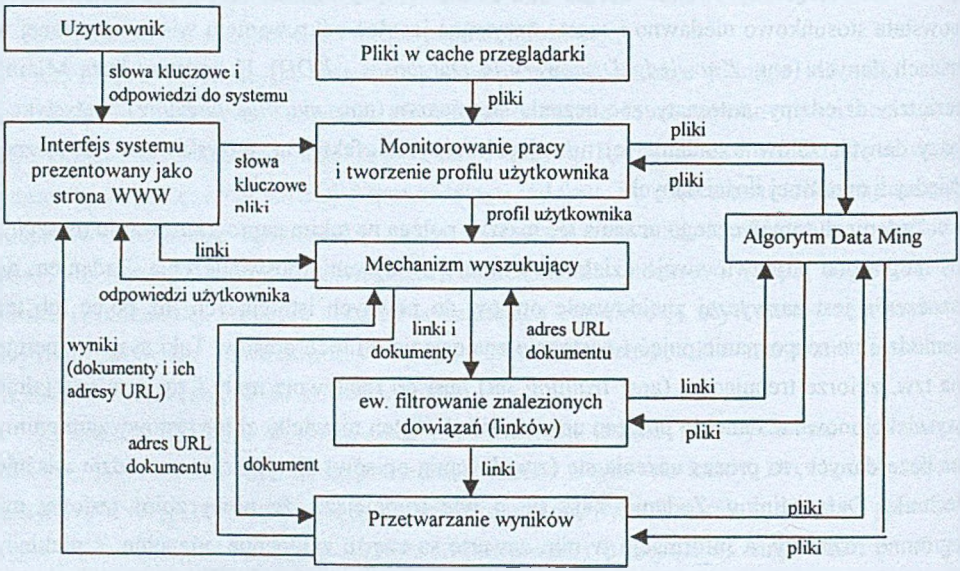
Zadanie automatycznego uczenia się maszyn polega na takim zaprogramowaniu maszyny, by mogła ona poprawić swoje działanie w miarę zdobywania doświadczenia. Zadaniem do nauczenia jest zazwyczaj znajdowanie opisów do pewnych istniejących już pojęć lub też samodzielne rozpoznanie pojęć i następnie znalezienie do nich opisów. Taki system operuje na tzw. zbiorze trenującym (ang. *training set*), jest on relatywnie mały i zawiera specjalnie wyselekcjonowane dane do procesu uczenia się. Jeśli ten niewielki zbiór testowy zamienimy na bazę danych, to proces uczenia się (znajdowania opisów) na tym zbiorze będzie właśnie techniką Data Mining. Zadanie staje się o tyle trudniejsze, że nowy zbiór testowy ma ogromne rozmiary, a informacje w nim zawarte są często zakłócone, niepełne, a niekiedy nawet sprzeczne. Do czyszczenia takich informacji używana są metody statystyczne [1].

Same techniki Data Mining przypominają metody statystyczne, ale jest pomiędzy nimi znacząca różnica, polegająca na tym, że w procesie Data Mining system towarzyszy użytkownikowi podczas generowania wyników uczenia się (tzw. hipotez). W metodach

statystycznych nie ma takiej możliwości. Data Mining zajmuje się wyszukiwaniem wszelakiego rodzaju zależności, powiązań i sekwencji czasowych oraz klasyfikowaniem.

Gdy potraktujemy światową sieć WWW jako wielką rozproszoną bazę danych, Data Mining jako inteligentne narzędzie oferuje nam wiele możliwości [2,6]. Użytkownik przeglądający sieć, gdy chce po raz kolejny znaleźć interesujące go materiały, szuka ich w tych samych lub podobnych miejscach w sieci, i szukając – wykonuje te same lub bardzo podobne czynności. Jeżeli zainteresowały go trzy strony dotyczące pewnego zagadnienia, to może też zainteresuje go czwarta, znajdująca się w podobnej lokalizacji? Dziedzina zajmująca się tymi zagadnieniami nazywana jest z angielskiego Web Mining.

By wykluczyć przypadek, należy przyjąć, że użytkownik uważa daną stronę za interesującą, gdy ogląda ją odpowiednio długo, lub że wykonał na niej jakąś akcję (zrobił do niej zakładkę, wydrukował). Wiedza (dane dla systemu Data Mining) na stronach WWW zapisana jest dwojako: w treści plików HTML oraz w powiązaniach pomiędzy tymi stronami (linkach) [3]. Jako materiał do nauki systemu można wziąć pod uwagę słowa kluczowe występujące na stronie (w tekście strony i/lub nagłówku) lub strony powiązane ze stroną, którą zainteresował się użytkownik. Można na bieżąco śledzić poczynania użytkownika (tzw. *on-line*) lub analizować strony obejrzane przez niego wcześniej, a które powinny znajdować się w cache przeglądarki internetowej (rys. 1).



Rys. 1. Budowa systemu poprawiającego pracę przeglądarki
Fig. 1. Construction of system, which improve efficient of web browser's work

3.1. Uczenie się na bazie słów kluczowych

System uczący się na bazie słów kluczowych analizuje zawartość stron internetowych przeglądanych przez użytkownika po to, by na ich podstawie dowiedzieć się jak najwięcej o zainteresowaniach użytkownika. Z treści stron wybierane są słowa lub całe frazy mające istotne znaczenie [2]. Przykładowa aplikacja to Syskill&Webert [7].

Algorytmem najczęściej wykorzystywanym przez tego typu systemy jest algorytm TFIDF (ang. *Term Frequency Inverse Document Frequency*), oparty na badaniu częstości występowania słów. Jest to algorytm (właściwie heurystyka) przypisujący wagi do każdego ze słów kluczowych występujących w dokumencie:

1) częstość występowania (ang. *Term Frequency*) danego słowa w_j w dokumencie, oznaczaną jako $TF(w_j)$,
2) liczbę dokumentów, w których dane słowo się pojawia (ang. *Document Frequency*), oznaczaną przez $DF(w_j)$.

Wartość wskaźnika TFIDF dla j-tego słowa wyliczana jest ze wzoru:

$$TFIDF(w_j) = TF(w_j) \cdot IDF(w_j) \tag{1}$$

gdzie $TF(w_j)$ jest miarą częstości występowania danego słowa w badanym dokumencie, $IDF(w_j)$ oznacza zaś liczbę dokumentów, w której nie pojawia się to słowo – miara ta jest logiczną odwrotnością miary $DF(w_j)$. Wartość $DF(w_j)$ obliczana jest ze wzoru:

$$IDF(w_j) = \log\left(\frac{|D|}{DF(w_j)}\right) \tag{2}$$

$|D|$ oznacza całkowitą liczbę dokumentów w zbiorze analizowanych dokumentów. Słowa z największymi wartościami $TFIDF$ są wybierane do nauki dla systemu [5].

3.2. Uczenie się na bazie powiązań pomiędzy stronami

Dwie strony uważane są za logicznie powiązane ze sobą, jeśli trzecia strona zawiera odnośniki (linki) do ich obu. Algorytm takiego uczenia się bazuje na tzw. informacji wzajemnej (ang. *mutual information*) [4,5]. Przykładowym systemem stosującym takie dane do uczenia się jest WebWatcher [4]. Analizowana jest tutaj struktura sieci wspomagana przez badanie zawartości poszczególnych stron. Słowa kluczowe potrzebne są do określenia zainteresowań użytkownika, a badanie stopnia powiązania stron internetowych pozwala określić, gdzie najlepiej szukać tej informacji.

4. Serwisy webowe

Nie tylko użytkownik ma problemy z siecią WWW, mają je również serwisy webowe. Największe to nadmiar zawartych w nich informacji, a co za tym idzie – niemożność dobrego jej przedstawienia. Często informacje błahe (naszym zdaniem) zajmują główne strony serwisów, podczas gdy informacji nam potrzebnych trzeba mozolnie szukać, nie zawsze zresztą z dobrym skutkiem. Być może pomogłoby zmodyfikowanie zawartości niektórych stron serwisu? Albo ich struktury? Może dodanie paru linków lub usunięcie kilku niepotrzebnych poprawi czytelność stron? Pozostaje nam stwierdzić, które ze stron są niepotrzebne, a które najbardziej oglądane.

Najprostszym narzędziem okazuje się analiza odwiedzin stron, przestawiana zazwyczaj w formie statystyki. Bardziej wyrafinowane narzędzia oferują firmy zajmujące się badaniem rynku internetowego: prowadzenie analizy odwiedzin stron łączą z pozyskiwaniem niektórych informacji od użytkownika (standardowo zapamiętywanych w logach serwisów webowych). Mogą to być np. adres IP komputera użytkownika, system operacyjny, którego używa, czas, jaki poświęcił na oglądanie strony. Jeśli to dla nas za mało, z pomocą może nam przyjść Data Mining.

Na podstawie informacji pozyskanych z logów serwera webowego, algorytmy Data Mining mogą spróbować znaleźć np. powiązania pomiędzy odwiedzanymi przez użytkownika stronami (jeśli wielu użytkowników odwiedza dwie konkretne strony, to może trzeba dodać linki pomiędzy nimi?), cechy wspólne dla użytkowników odwiedzających określone strony po to, by następnie stworzyć specjalne profile użytkowników – zgodne z ich zainteresowaniami – na serwerach WWW. Istnieje możliwość prześledzenia ścieżek, którymi wędrują użytkownicy po to, by ewentualnie zmodyfikować strukturę stron WWW na tym serwisie. Można linki do pewnych często odwiedzanych stron umieścić np. na stronie głównej. Wyjątek stanowią serwisy, których zawartość wymusza określoną kolejność oglądania stron: usługi należące do e-commerce, serwisy wymagające subskrypcji za swoje usługi (trzeba się zalogować) oraz np. serwisy pozwalające założyć konto poczty elektronicznej.

5. Podsumowanie

Inteligentne aplikacje, wspierające korzystanie z serwisów webowych zarówno od strony użytkownika (sprzężone z przeglądarkami stron WWW), jak i na serwerach stron internetowych, umożliwiają zwiększenie efektywności pracy i pozwalają nam nieco

zapanować nad ogromem informacji znajdującej się w sieci. Przedstawione powyżej rozwiązania mają niewątpliwie swoje zalety, ale mają też i wady. Pierwszym i najważniejszym ich mankamentem jest to, że nie są jeszcze w powszechnym użyciu. Może to świadczyć o ich brakach, ale mogą być też inne przyczyny, bardziej ludzkiej natury.

Wzrasta liczba internautów, którzy są przeciwni wszelakim formom inwigilacji w sieci. Również wtedy, jeśli jest to jedynie pozyskiwanie niewiele znaczących informacji o ich pracy. Zbyt wiele informacji zebrać nie można, chociażby ze względu na obowiązujące prawo chroniące prywatność. Pojawiają się już systemy, które na podstawie ogólnodostępnych informacji o użytkownikach sieci usiłują domyśleć się np. płci, wieku, wykształcenia, ale daleko im do doskonałości.

Przy tworzeniu systemów wspomagających przeglądanie przez użytkownika sieci WWW nie wolno zapominać o jednym, najważniejszym fakcie: inteligentne systemy przeszukiwania mogą jedynie przedstawiać swoje sugestie – to człowiek decyduje, co chce znaleźć, jakie strony odwiedzić i kiedy zakończyć poszukiwania.

LITERATURA

1. Cichosz P.: Systemy uczące się. Wydawnictwa Naukowo-Techniczne,Warszawa 2000.
2. Drummond Ch., Ionescu D., Holte R.: A Learning Agent that Assists the Browsing Software Libraries. Computer Science Department, University of Ottawa, Technical Report TR-95-12.
3. Holte R., Drummond Ch.: A Learning Apprentice For Browsing. Proceedings of the 1994 AAAI Spring Symposium on Software Agents, Stanfrod, AAAI Press.
4. Joachims T., Mitchell T., Freitag D., Armstrong R.: WebWatcher: Machine Learning and Hypertext. Fachgruppentreffen Maschinelles Lernen, Dortmund, Niemcy.
5. de Kroon H.C.M., Mitchell T.M., Kerckhoffs E.J.H.: Improving Learning Accurancy in Information Filtering. Internatonal Conference on Machine Learning – Workshop on Machine Learning Meets HCI (ICML–96).
6. Leong H., Kapur S., de Vel O.: Text Summarization for Knowledge Filtering. Agents in Distributed Heterogeneous Environments. Department of Computer Science James Cook Uniwersity of North Queensland, Australia, Technical Report.
7. Pazzani M., Muramatsu J., Billsus D.: Syskill&Webert: Identyfying interestning Web Sites. AAAI Spring Symposium, Stanford.

Recenzent: Dr inż. Marcin Gorawski

Wpłynęło do Redakcji 5 kwietnia 2002 r.

Abstract

Nowadays World Wide Web is the largest library that provides us a lot of information. Often it is difficult to find right piece of information. There is a need to build software which helps a user find it. Application that learns user's interest, creates special user's profile and then suggests hyperlink to pages about interesting topics can help user obtain significant and proper information. Figure 1 shows functional modules in an intelligent browsing system. Web services wrestle with a lot of problems too. They have too much information, which is not presented in the right way, and is located in wrong places or sometimes even hidden. Web services need software, which effective assist presentation of information. On these services, user's profiles can be composed dynamically after user's interest recognition. We can also modify the source or sequence of web pages. This paper approaches some solutions, which were created in order to make easier browsing easier.