STUDIA INFORMATICA 2004
Volume 25 Number 3 (59)

Karol KOZAK
Max Planck Society, ZIM Centre for Information Management

METODY ZWIĘKSZAJĄCE PRECYZYJNOŚĆ WYSZUKIWANIA INFORMACJI – AUTOMATYCZNA KATEGORYZACJA

Streszczenie. Niniejszy artykuł poświęcony jest metodom organizacji dokumentów uzyskanych w wyniku wyszukiwania i ich zastosowania w aplikacjach. Opisana metoda polega na automatycznej kategoryzacji dokumentów, a przykład zastosowania tej metody w systemach zawierających w swoim repozytorium dokumenty pochodzi z dziedziny medycyny – psychiatrii.

Słowa kluczowe: lingwistyka komputerowa, eksploracja informacji, medyczne bazy danych, klasyfikacja dokumentów, metadata publikacji, archiwalne bazy danych.

METHODS FOR INCREASING PRECISION IN FIND INFORMATIONS – AUTOMATIC CATEGORIZATION

Summary. In this article are described methods and their implementation to the systems for organize documents from search results. Below is described approach that automatically categorize documents. This approach was implemented to systems which contain and are repository for scientific documents from domain medicine - psychiatry.

Keywords: information retrieval, text mining, medical informatics, document classification, archival sources, metadata documents.

1. Wstęp

Źródła informacji zawierające dokumenty, takie jak publikacje naukowe pochodzące z różnych czasopism, są ogromnym repozytorium, które stale wzrasta. Zwiększanie się ilości danych w tych repozytoriach jest efektem dokładania dokumentów, które są aktualnie tworzone, także poprzez proces budowania elektronicznych wersji dokumentów z archiwalnych

źródeł w dziedzinie medycyny. Poprzez tworzenia danych pochodzących z archiwalnych źródeł zostaje ułatwiony dostęp do nich. Informacje w nich zawarte w istotny sposób wpływają na rozwój aktualnej nauki. Przykładem takiego repozytorium, które stale zwiększa swoją wielkość, jest system „PUBMED" [10] zawierający 14 milionów medycznych publikacji. W artykule został opisany sposób automatycznej kategoryzacji w tematyczne grupy dokumentów w kolekcji, uzyskanej w wyniku wyszukiwania. Poniższe metody automatycznej kategoryzacji składają się z dwóch głównych elementów: modelu zapytania (ang. *query model*) i modelu terminologicznego (ang. *terminology model*), które istotnie wpływają na pracę kategoryzatora i pozwalają zautomatyzować proces kategoryzacji. W celu sprawdzenia praktycznego działania i uzyskania wyników porównawczych sposobu został on zastosowany w systemie e-doc [12] (elektroniczny dokument) oraz w myDM [11] (*my Document Manager*). Systemy te są repozytorium dokumentów z dziedziny medycyny: psychiatrii.

Poprzez przyporządkowanie dokumentów do poszczególnych kategorii jesteśmy w stanie w szybkim czasie znaleźć dokumenty, odpowiadające naszemu zapytaniu. Innymi sposobami, które automatycznie organizują dokumenty, jest szeregowanie zależne (ang. *relevance-ranking*) oraz systemy grupowania (ang. *clustering system*). Systemy oparte na szeregowaniu zależnym [2] tworzą uporządkowaną listę wyniku szukania. Uporządkowanie dokumentów bazuje na pomiarze oddzielnie każdego dokumentu i dopasowaniu go jak najlepiej do zadanego zapytania. Systemy bazujące na grupowaniu (ang. *clustering*) budują grupy dokumentów, opierając się na zależnościach, relacjach pomiędzy nimi (Allen, Obry et. Al. 1993; Hearst and Pederson 1996)(Saham, Zusufali et. al. 1998) [1]. W celu określenia zależności pomiędzy dokumentami systemy grupowania wymagają podobieństwa metrycznego. Przykładem takiej zależności jest liczba słów, które najczęściej występują w dokumentach.

Zanim zastosujemy kategoryzację, musimy najpierw ukazać w formie wektora kolekcję dokumentów uzyskaną w wyniku wyszukiwania. Wektorem dokumentu jest wielowymiarowa tablica członów. Do tych członów nie są zaliczane słowa, takie jak spójniki, przysłówki. Przykładowa forma dokumentu wektora wielowymiarowej tablicy członów:

Dokument 1 człon1 człon2 człon3 ...
Dokument 2 człon1 człon2 człon3 ...
Dokument 3 człon1 człon2 człon3 ...
....

W tym przypadku został zastosowany algorytm „Vector Space" [3] (Salton, Wond, et. al. 1975; Salton and McGill 1983; Salton 1989). W algorytmie "Vector Space" w celu zbadania ważności badanego członu w kolekcji jest stosowany wzór (1):

$$W_i = tf * log\ N/\ df_i \tag{1}$$

gdzie: W_i – oznacza wielkość opisującą ważność członu w kolekcji dokumentu, N – całkowita liczba dokumentów w kolekcji, tf – częstotliwość występowania członu w dokumencie, df_i – częstotliwość występowania członu w całej kolekcji. I-ty element w wektorze dokumentu reprezentuje wartość *i*-tego członu w tym dokumencie. Dokumenty są przedstawione w postaci wielowymiarowej tablicy członów, dzięki temu łatwiej jest poddać je eksploracji.

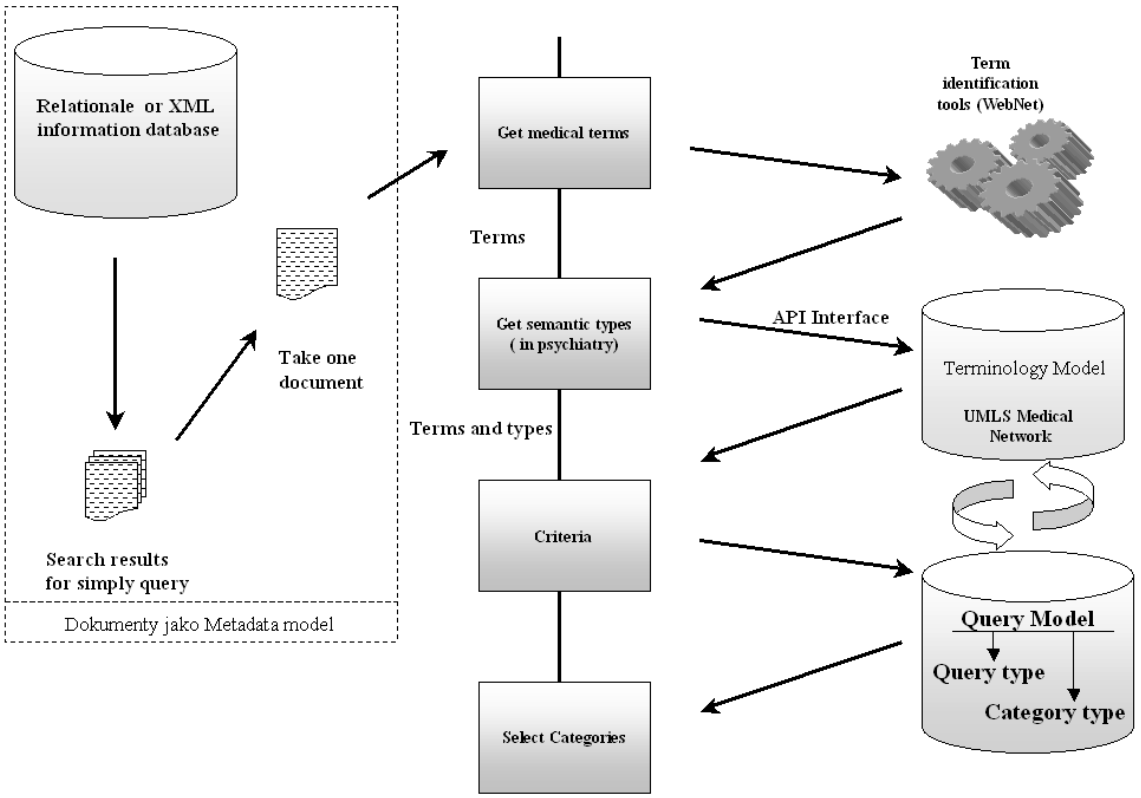
2. Specyfikacja systemu

2.1. Model metadanych dokumentów

Dokumenty utrzymywane w repozytorium posiadają pewną strukturę. Taką strukturę w przypadku dokumentów naukowych nazywamy modelem metadanych. Dla opisywanego systemu w dziedzinie medycyny- psychiatrii został stworzony własny model. Nie można było skorzystać ze standardowego modelu publikacji, gdyż źródłami są też dokumenty archiwalne od roku 1950. Model składa się z zagnieżdżonych elementów, które najlepiej jest ukazać za pomocą schematu XML. Jest on rozszerzeniem, modyfikacją modelu metadanych „dublin core" [13]. Model ten zaimplementowany został w systemach: e-doc i myDM. Do głównych elementów modelu należy: title, creator (role), content type, abstract, volume (spage, epage), language, enduser, publish date, date modified, MeSH (Medline subject headline) [14].

2.2. Komponenty

Do zbudowania systemu z automatyczną kategoryzacją niezbędne będzie utworzenie takich elementów, jak: model zapytania (ang. *query model*) i model terminologiczny (ang. *terminology model*). Problem, który trzeba rozwiązać to znalezienie sposobu na automatyczne tworzenie modelu zapytania (QM). Badania zostały przeprowadzone na podstawie publikacji z dziedziny psychiatrii przy wykorzystaniu modelu terminologicznego (TM) stworzonego przez *National Library of Medicine* (NLM) [4] (50 000 medycznych zagadnień). Dostęp do modelu terminologicznego uzyskujemy poprzez zastosowanie API (*Aplication Programing Interface*) dla języka Java [6].
Rys. 1 przedstawia nam przebieg działania opisywanego sposobu z użyciem wszystkich niezbędnych elementów.



Rys. 1. Schemat procesu automatycznej kategoryzacji z użyciem modelu terminologicznego – UMLS
Fig. 1. Schema for automatic categorization pocess with terminology model – UMLS

2.2.1. Model zapytań

Do zorganizowania dokumentów w kategorie uzyskane w wyniku wyszukiwania, które odpowiadają postawionemu zapytaniu, system wymaga wiedzy, jakiego rodzaju zapytania są stawiane w danej dziedzinie. Stworzony model zapytań dla repozytorium psychiatrii składa się z dwóch części: typu zapytania (ang. *query type*) oraz typu kategorii (ang. *category type*). W każdej dziedzinie nauki, jeśli będziemy chcieli skorzystać z wyżej opisanej metody kategoryzacji, musimy ręcznie lub automatycznie stworzyć model zapytań (QM). Model zapytań w tym przypadku został automatycznie zbudowany za pomocą analizy leksykalnej członów medycznych w dokumentach. Wynikiem tego procesu jest utworzenie typów zapytań (ang. *query types*), co w tym przypadku (w medycynie) było efektem powstania między innymi takich typów, jak: prevention, diagnosis, treatment, prognosis. Analiza leksykalna została przeprowadzona na zapytaniach stawianych podczas przeszukiwania informacji w czasopismach naukowych z zakresu psychiatrii (PUBMED – *journals*) [10].

Typ kategorii – dla każdego rodzaju zapytania system potrzebuje także jego abstrakcyjnego znaczenia w celu zbudowania tematu lub kategorii odpowiadającego grupie uzyskanej

w rezultacie wyszukiwania. Model zapytania (QM) w systemie e-doc i myDM przyjął 9 takich typów dla kategorii: problems, symptoms, preventive- actions, risk factors, diagnoses, tests, treatments, prognoses and prognostic- indicators. Każdy typ zapytania w QM jest przyłączony do typu kategorii określającej rodzaj kategorii generowanej w systemie.

2.2.2. Model terminologiczny

W celu określenia odpowiedniego opisu kategorii dla danej grupy dokumentów system powinien posiadać informacje, który opis kategorii jest odpowiedni dla podanego typu kategorii. Terminologiczny model (TM) dostarcza informacji poprzez połączenie poszczególnych członów (pojedynczych słów, akronimów, wielowyrazowych członów) wraz z ich nadrzędnymi znaczeniami nazywanymi typami semantycznymi. Takim terminologicznym modelem jest właśnie system UMLS (*Unified Medical Language System*) [5]. UMLS wiąże każdy człon do przynajmniej jednego typu semantycznego w sieci semantycznej. Indywidualne, specyficzne człony mogą otrzymać opis kategorii, jeśli ich typ semantyczny jest połączony z żądanym typem kategorii. Model zapytań (QM) jest ściśle połączony z modelem terminologicznym, gdyż każdy typ kategorii jest określony przez listę semantycznych typów pochodzących właśnie z UMLS. Człony medyczne w TM mogą zmieniać się dynamicznie, mogą dochodzić nowe typy semantyczne. Dlatego QM powinien być tak zbudowany, żeby mógł współpracować z TM, ale w sposób całkowicie niezależny od TM. Każda zmiana w TM nie powinna powodować naruszenia prawidłowego działania QM.

2.3. Proces kategoryzacji

Aby móc przedstawić, jak działa algorytm kategoryzacji, należy wcześniej opisać elementy, które wchodzą w jego skład, czyli QM, TM i model metadanych. Aby sprawdzić opisany sposób kategoryzacji, wszystkie elementy zostały zastosowane w systemach e-doc i myDM. Kategoryzacja zaczyna się od momentu uzyskania kolekcji dokumentów w wyniku wyszukiwania dla pewnego zapytania. Kategoryzator musi najpierw usunąć z kolekcji dokumentów elementy nie pasujące do listy potencjalnych kategorii z QM. Niektóre znaczenia semantyczne słów kluczowych w kolekcji dokumentów nie odpowiadają do zapytaniu użytkownika. Dla każdego dokumentu kategoryzator sprawdza słowa kluczowe zgodnie z kryteriami odpowiadającymi typowi zapytania z QM. Sprawdza dla każdego słowa typ semantyczny (ang. *semantic type*) z TM i porównuje go z listą zaakceptowanych typów semantycznych w kryteriach kategoryzacji. Gdy dane słowa kluczowe spełniają wszystkie kryteria kategoryzacji, kategoryzator dodaje dokument do określonej kategorii opisanej tym kluczem. Jeśli taka kategoria jeszcze nie istnieje, dodaje ją jako nową. Jeśli już istniała, dodaje dokument do kategorii o tej samej nazwie co sprawdzane słowo kluczowe. Według

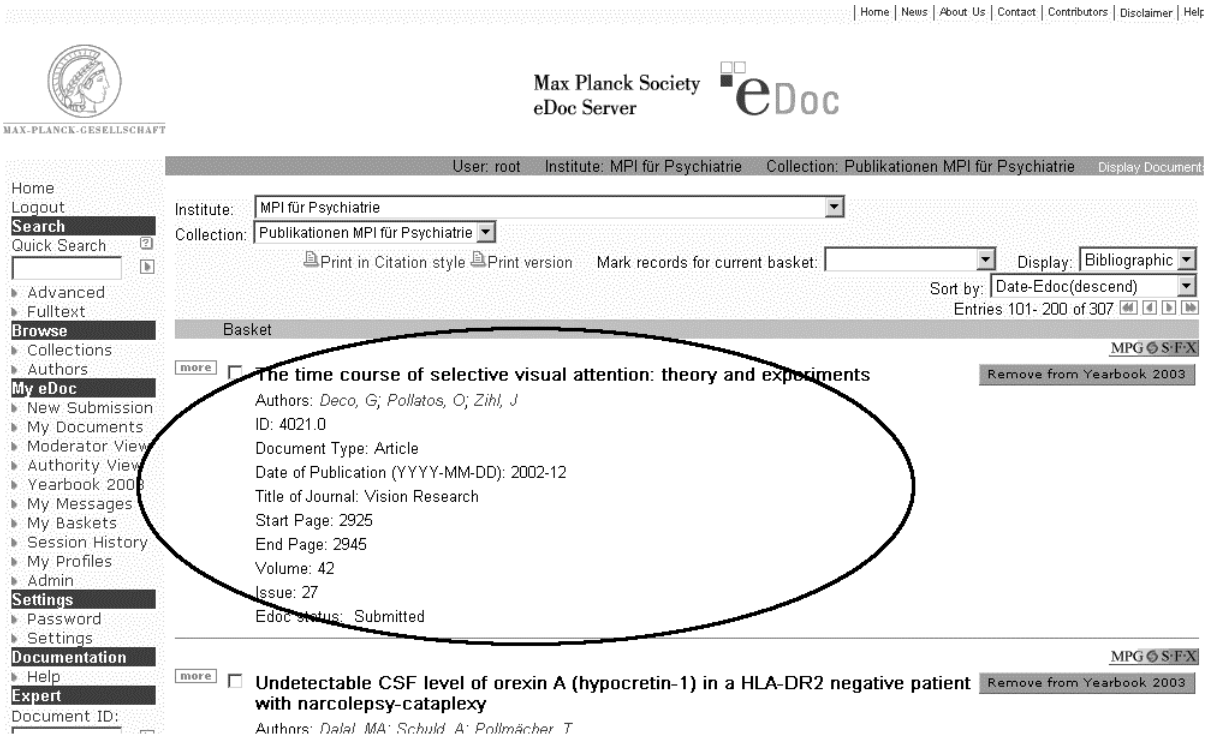
takiego schematu są sprawdzane wszystkie słowa kluczowe w dokumentach. Klucze nie spełniające kryterium są po prostu ignorowane.

3. Implementacja algorytmu

W celu sprawdzenia skuteczności działania oraz uzyskania wyników opisany sposób na automatyczną organizacje dokumentów zaimplementowany został do dwóch projektów: e-doc i myDM.

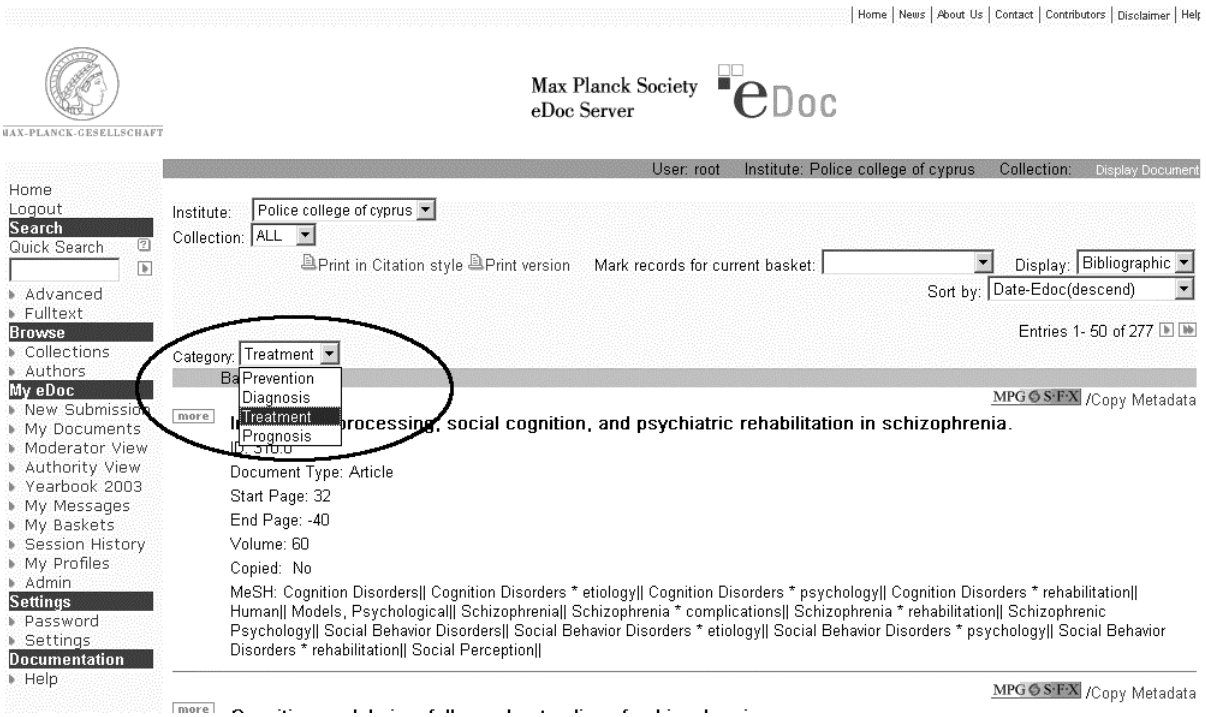
3.1. E-doc

E-doc jest to system o strukturze wielowarstwowej klient – serwer. Jest to system, gdzie w relacyjnej bazie danych umieszczone są publikacje naukowe z dziedziny medycyny – psychiatrii.



Rys .2. E-doc – interfejs użytkownika umożliwiający przeglądanie kolekcji dokumentów uzyskanych w wyniku wyszukiwania

Fig. 2. E-doc – user interface for browsing documents from search result



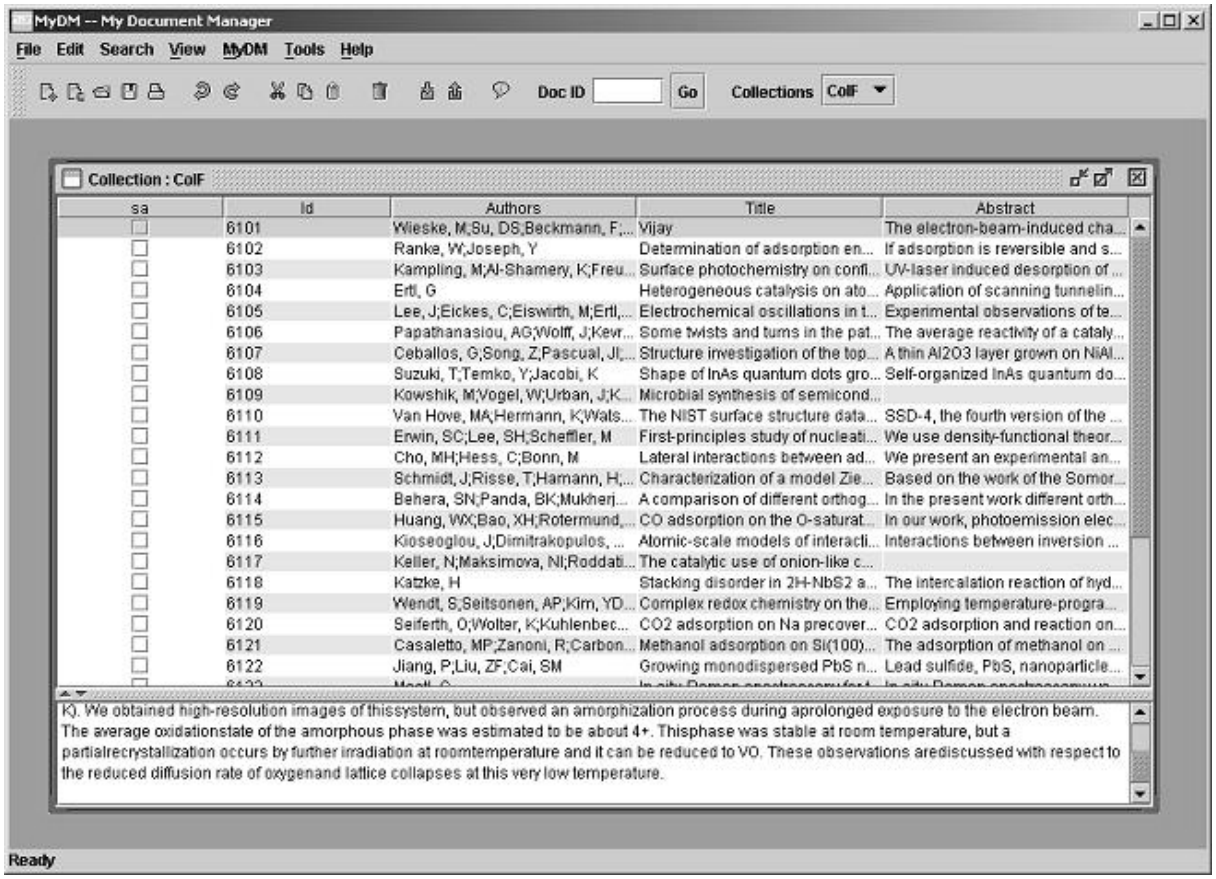
Rys. 3. E-doc - interfejs użytkownika umożliwiający przeglądanie dokumentów w poszczególnych kategoriach po zastosowaniu kategoryzacji
Fig. 3. E-doc – user interface for browsing documents in selected category from search result after apply categorisation

Poprzez zapytania do bazy danych uzyskaliśmy kolekcje dokumentów (rys. 2). Dalej przez udostępniony mechanizm możemy przeprowadzić kategoryzacje dokumentów z tej kolekcji (rys. 3).

Projekt ten jest wykonany na bazie języka programowania Embedded Perl ze względu na jego dobre właściwości pracy z tekstem. Interfejsem użytkownika jest www. Baza danych to PostgresSQL [8], a jako web serwer został zastosowany serwer Apache [7].

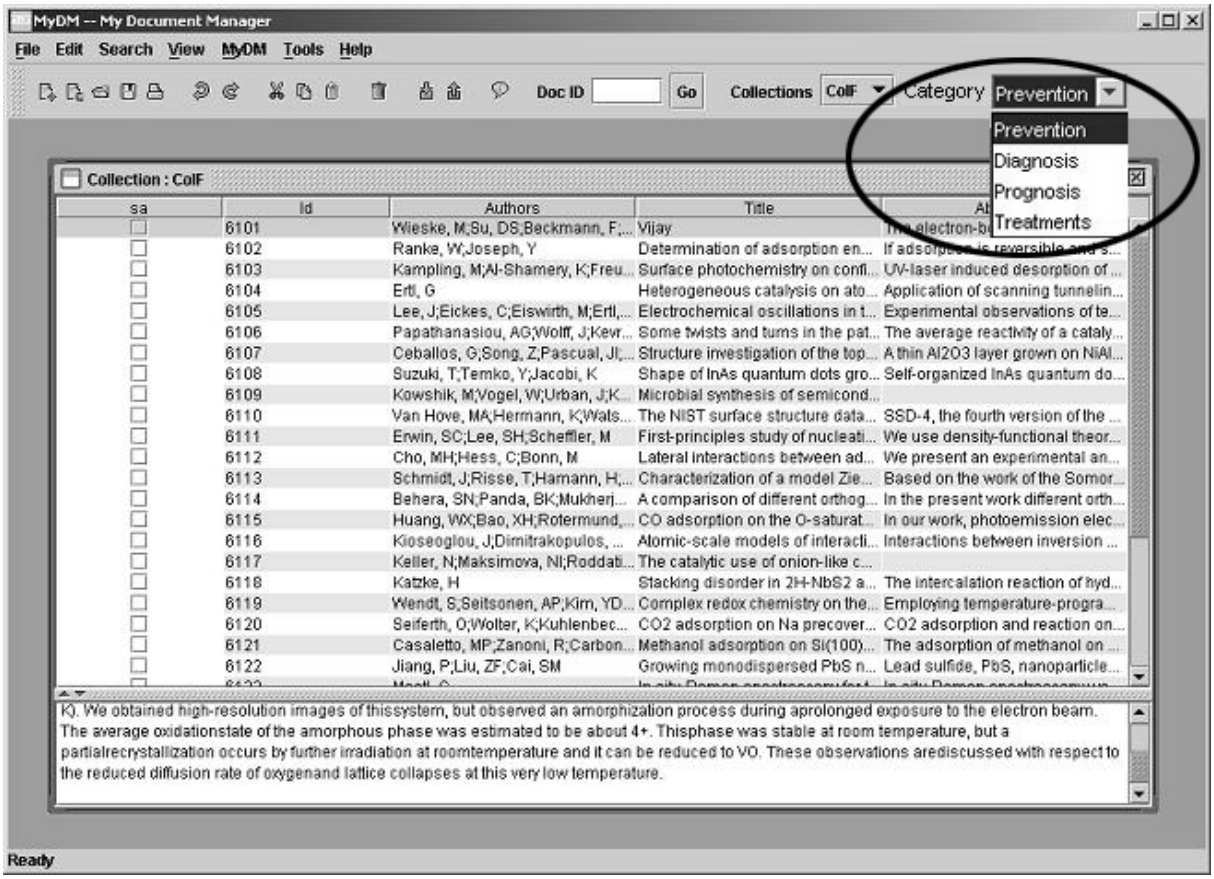
3.2. MyDM

MyDM jest podobnym projektem do e-doc, z tą różnicą że jest to oprogramowanie jednowarstwowe, instalowane na maszynie klienta. Tutaj baza danych także zawiera dokumenty z dziedziny psychiatrii. W myDM został także zaimplementowany algorytm kategoryzacji (rys. 4, rys. 5). Projekt ten jest wykonany kompleksowo w języku Java [6], gdzie interfejsem użytkownika jest „Java Swing”. Baza danych to HSQL [9].



Rys. 4. MyDM – interfejs użytkownika umożliwiający przeglądanie dokumentów uzyskanych w wyniku wyszukiwania

Fig. 4. MyDM – user interface for browsing documents from search result



Rys. 5. MyDM – interfejs użytkownika umożliwiający przeglądanie dokumentów w poszczególnych kategoriach po zastosowaniu kategoryzacji

Fig. 5. MyDM – user interface for browsing documents in selected category from search result after apply categorisation

4. Ocena

Po zbadaniu działania algorytmów w projektach e-doc i myDM można było wyciągnąć wnioski. Rezultaty organizacji dokumentów z wykorzystaniem kategoryzacji, jakie zostały uzyskane z projektów, wypadły pozytywnie w porównaniu do innych narzędzi (*clustering tool*, *ranking tool*). Aby można było porównać wyniki, zastosowano to samo zapytanie „*pathology deases*" do wszystkich sposobów dla tej samej bazy danych. Pod względem stabilności systemu (ang. *performance*) e-doc wypadł trochę gorzej niż myDM. Powodem tego była budowa myDM oparta na języku Java, dzięki czemu można było używać bezpośrednio API dla TM UMLS. W e-doc potrzeba było użyć specjalnych protokołów, dzięki którym można było uzyskać połączenie z API dla UMLS. Po zaimplementowaniu kategoryzacji w systemie e-doc użytkownicy byli usatysfakcjonowani z uzyskanych rezultatów wyszukiwania. Do tej pory system pracował bez elementów kategoryzacji. Po wprowadzeniu kategoryzacji wielu użytkowników uznało kategoryzację za poprawną i po-

mocną w przeglądaniu dużej ilości dokumentów w kolekcji uzyskanych w wyniku wyszukiwania.

5. Podsumowanie

W artykule został zaprezentowany sposób, dzięki któremu istnieją możliwości tworzenia systemów z automatyczną kategoryzacją wyników wyszukiwania. Budując takie systemy, jesteśmy w stanie stworzyć automatycznie nazwy kategorii. Zbadano, jak działa organizacja dokumentów i czy jest prawidłowa w odniesieniu do typu zadawanego zapytania. Jakie dokumenty, publikacje z różnych lat z dziedziny psychiatrii można powiązać pod względem znaczenia semantycznego? Stosowanie takiego sposobu do systemów wolnego dostępu (ang. *open access*), jakim jest e-doc, pomoże użytkownikowi w wyszukiwaniu informacji. Użytkownicy (lekarze, naukowcy, studenci) mogą dzięki takim systemom uzyskać łatwiej informacje z dokumentów archiwalnych.

LITERATURA

1. Sahami M, Yusufali S, Baldonado M.Q: Service for Organizing Network Information Autonomously. Digital Libraries 98: Proceedings of the Third ACM Conference on Digital Libraries, Pittsburgh, PA, USA 1998.
2. Buckley C, Salton G, Allan J: The effect of adding relevance information in a relevance feedback environment. SIGIR ’94. Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Springer-Verlag, Berlin,1994.
3. Salton G, Wong A, Yang CS: A vector space model for automatic indexing. Communications of the ACM 18: 1975, s. 613-620..
4. NLM. National Library of Medicine. http://www.nlm.nih.gov/. 2003.
5. UMLS. Unified Medical Language System Knowledge Source Server. [Online] http://umlsks1.nlm.nih.gov. 2003.
6. Java technology. [Online] http://www.java.sun.com. 2003.
7. Apache. Open-source HTTP server. http://httpd.apache.org/. 2003.
8. Postgres Object-relational database management system.http://www.postgres.org. 2003.
9. HSQL. Open source database [Online] http://hsqldb.sourceforge.net/. 2003.
10. NLM. Welcome to PubMed.[Online] http://www.ncbi.nlm.nih.gov/entrez/. 2003.
11. My Document Manager. [Online] http://mydm.sourceforge.net/. 2003.

12. Heinz Nixdorf Center for Information Management in the Max Planck Society. [Online] http://www.edoc.mpg.de. 2003.
13. Dublin Core Metadata Initiative. [Online] http://dublincore.org/. 2003.
14. MeSH. Medline Subject Headings [Online] http://www.nlm.nih.gov/pubs/factsheet . 2003.

Recenzent: Dr inż. Marcin Gorawski

Wpłynęło do Redakcji 18 maja 2004 r.

Abstract

The amount of information coming from public scientific sources is growing rapidly. People that searching information in big amount documents are not able to find concrete response in quick time. For help people in this problem we need methods and their implementation into systems for organize documents from search results. It was developed approach that automatically categorize documents. This method implemented to system create from results search reasonable document categories and assign documents to categories appropriately. Groups documents in category are organize meaningful because correspond to the user's query. Every query is analyze uses knowledge of important kind of queries and a model of the domain terminology. I prepared two tools E-doc and MyDM that implements this approach for the domain medicine-psychiatry. This projects are repositories where the amount of information in the primary psychiatry literature alone is overwhelming. After apply categorize both system summarize the documents from search by organizing them into an intuitive and useful categories, thus helping users to gain quick and easy access to important medical information.

Adres

Karol KOZAK: Osiedlowa 11/13, 26-600 Radom (lub Gueterbahnhofstr 10/501, 01067 Dresden, Germany), karkoz1@gmx.de .