XXII Seminarium

ZASTOSOWANIE KOMPUTERÓW W NAUCE I TECHNICE' 2012
Oddział Gdański PTETiS
Referat nr 13

METODY WPROWADZANIA INFORMACJI DO MOBILNYCH DOKUMENTÓW INTERAKTYWNYCH OPARTE NA IDENTYFIKACJI PODOBNYCH TREŚCI

Adam Łukasz KACZMAREK[1]

1. Politechnika Gdańska; Wydział Elektroniki, Telekomunikacji i Informatyki
tel: 58 347 13 78 fax: 58 347 22 22 e-mail: adam.l.kaczmarek@eti.pg.gda.pl

Streszczenie: Artykuł dotyczy nowatorskiej architektury dokumentu pozwalającej na szybsze wprowadzanie informacji i wydajniejsze ich przetwarzanie. Dokumenty opracowywane zgodnie z przedstawioną architekturą charakteryzują się tym, że są wykonywalne, mobilne, interaktywne oraz inteligentne. Dokumenty te aktywnie współpracują z użytkownikiem podczas wprowadzania treści oraz są w stanie automatycznie przemieszczać się w Internecie. Ponadto dokumenty, oprócz wizualnego przedstawienia treści w nich zawartych, posiadają również funkcjonalność realizującą te treści. Artykuł przedstawia możliwości zastosowania metod wykrywania plagiatów oraz metod wyszukiwania informacji do wspomagania użytkowników w tworzeniu dokumentów. W szczególności przedstawione jest wykorzystanie do wypełniania dokumentów metody klasteryzacji kierunkowej służącej do wspomagania wyszukiwania informacji.

Słowa kluczowe: dokumenty elektroniczne, wyszukiwanie informacji

1. WPROWADZENIE

Osoby korzystające w swojej pracy z komputerów znaczną część czasu poświęcają na tworzenie dokumentów. Usprawnienie procesu sporządzania dokumentacji zwiększyłoby wydajność pracowników i zaoszczędziło ich czas. Niniejszy artykuł dotyczy tematu nowoczesnych metod tworzenia dokumentów opartych na architekturze dokumentu zwiększającej możliwości dotychczas stosowanych formatów dokumentów.

Metody wydajniejszego tworzenia dokumentów dzięki lepszej ich architekturze opracowywane są w ramach projektu MENAID (MEtody i NArzędzia Inżynierii Dokumentu przyszłości) realizowanego na wydziale Elektroniki, Telekomunikacji i Informatyki Politechniki Gdańskiej [1]. Niniejszy artykuł poświęcony jest wykorzystaniu metody klasteryzacji kierunkowej [2] oraz metod wykrywania plagiatów [3] do wspierania użytkowników w tworzeniu dokumentów z wykorzystaniem architektury opracowywanej w projekcie MENAID.

Struktura artykułu opracowana została w taki sposób, że po pierwszej części będącej wstępem, przedstawione są główne założenia typu dokumentów tworzonych w projekcie MENAID. Następnie w rozdziale trzecim przedstawiona jest warstwowa architektura dokumentu. Rozdział czwarty poświęcony jest metodom wykrywania plagiatów. Rozdział piąty opisuje metodę klasteryzacji kierunkowej. W rozdziale szóstym przedstawiony jest sposób wykorzystania metod wykrywania plagiatów i metody klasteryzacji kierunkowej do wydajniejszego tworzenia dokumentów. Podsumowanie przedstawione jest w rozdziale siódmym.

2. FUNKCJONALNOŚĆ DOKUMENTU W PROJEKCIE MENAID

Dokumenty opracowywane w ramach projektu MENAID charakteryzują się tym, że są wykonywalne, mobilne, interaktywne i inteligentne.

2.1. Wykonywalność

W powszechnie używanych dokumentach zawarte jest jedynie przedstawienie graficzne informacji, które są treścią tych dokumentów. Informacje te mogą zostać odczytane, a następnie wykorzystane, jednak same dokumenty nie udostępniają funkcjonalności pozwalającej na przetwarzanie informacji w nich zawartych. Na przykład w publikacjach naukowych znajdują się opisy wzorów matematycznych oraz algorytmów, które wykorzystać można tylko w taki sposób, że na podstawie opisanych w dokumencie informacji przeprowadzana jest realizacja obliczeń. Jednak nie jest możliwe wykonanie obliczeń bezpośrednio za pomocą dokumentu.

W dokumencie elektronicznym zawarta może być oprócz opisu graficznego wzorów i algorytmów również ich funkcjonalność. Można powiedzieć, że dokument zawierałby program realizujący przedstawione graficznie informacje. Użytkownik, w szczególności recenzent publikacji, widząc pewną formułę mógłby wprowadzić własne dane i zweryfikować uzyskane wyniki. Funkcje przeprowadzające te obliczenia stanowiłyby integralną część dokumentu. Takiego rodzaju dokumenty nazywane są wykonywalnymi.

W celu tworzenia tego rodzaju dokumentów powstała pod patronatem wydawnictwa Elsevier inicjatywa Executable Paper Grand Challenge [4]. Jej celem jest opracowanie standardów tworzenia dokumentów wykonywalnych. Szczególnie istotne jest to, aby funkcjonalność dokumentu była niezależna od platformy sprzętowej oraz systemu operacyjnego, na którym dokument jest otwierany. Jednym ze sposobów tworzenia takich

Recenzent: Prof. dr hab. inż. Bogdan Wiszniewski – Wydział Elektroniki, Telekomunikacji i Informatyki Politechnika Gdańska

dokumentów jest użycie architektury IODA (Interactive Open Document Architecture) [5]. Charakteryzuje się ona tym, że dokumenty zgodne z tą architekturą są wielowarstwowe. Ponadto występuje warstwa, która służy do wykonywania zawartej w dokumencie funkcjonalności.

2.2. Mobilność

Oprócz utworzenia i wypełnienia dokumentów konieczne jest dostarczenie ich do osób, dla których dokumenty te są przeznaczone. Dokumenty są regularnie przesyłane w postaci załączników do poczty elektronicznej. Użytkownicy ręcznie wybierają adresatów i przesyłają im pliki. Dokumenty jednak mogłyby same automatycznie przesyłać się do odpowiednich osób w momencie, gdy zostają wypełnione. Na tym polega mobilność dokumentu.

Uwzględniana jest możliwość występowania sytuacji, w której wypełnienie dokumentu wymaga ingerencji wielu osób. Dotyczy to na przykład dokumentów sądowych, które są uzupełniane przez różne osoby lub list ocen studentów uzupełnianych przez wiele osób prowadzących zajęcia. Dokument może mieć w sobie zawartą sekwencję osób, które powinny go przetwarzać. Jest również możliwe to, że dokument utworzy swoje kopie, które zostają jednocześnie przesłane do wielu różnych osób, a następnie po wprowadzeniu danych przez te osoby, uzyskane informacje zostają scalone w jednym dokumencie. Tego rodzaju przetwarzanie dokumentów możliwe jest dzięki zastosowaniu architektury MIND (Mobile Interactive Documents) [6][7]. W architekturze tej dokumenty posiadają określone przepływy (ang. workflow), zgodnie z którymi dokumenty migrują do różnych użytkowników.

2.3. Interaktywność

Interaktywność dokumentu polega na tym, że współpracuje on z użytkownikiem podczas wypełniania. Przede wszystkim dokument po dostarczeniu do użytkownika podaje komunikat, że wymagana jest aktywność użytkownika polegająca na dodaniu do dokumentu informacji. Ponadto podczas wprowadzania informacji użytkownikowi prezentowane są liczne sugestie i podpowiedzi ułatwiające wypełnianie dokumentu. Dokument udostępnia użytkownikom funkcjonalność automatycznego uzupełniania tekstu na podstawie wpisywanych przez użytkownika liter. Ponadto, w przypadku gdy dokument ma postać formularza, w poszczególnych polach wypełnianych przez użytkownika pojawiają się proponowane wyrażenia. Dokument udostępnia również wsparcie w wypełnianiu na urządzeniach mobilnych.

2.4. Inteligencja

Inteligencja dokumentu polega na tym, że dokument ten, podobnie jak agent w systemach wieloagentowych [7] ma pewien cel, który realizuje. Tym celem jest wypełnienie dokumentu w wyniku zebrania wszystkich wymaganych informacji. Dokument może realizować ten cel na różne sposoby. Może on się przesłać do pewnej osoby posiadającej te informacje, a jeśli ta osoba jest niedostępna przez dłuższy czas lub nie udziela odpowiednich informacji, dokument może się przesłać do innej osoby. Dokument może posiadać informacje o innych osobach mogących go wypełnić i do tych właśnie przesłać się w celu wypełnienia.

Możliwe jest również to, że dokument podzieli się na części i jedynie te części dokumentu zostaną przesłane do odpowiednich użytkowników. Gdy użytkownicy wprowadzą informacje dokument jest scalany w całość. Taki sposób przetwarzania dokumentu jest szczególnie istotny wtedy, gdy występują użytkownicy, którzy nie są uprawnieni do przeglądania całego dokumentu. Użytkownikom takim przesyłane są tylko te części dokumentu, do których powinni mieć dostęp.

Ponadto inteligentny dokument potrafi wypełniać się samodzielnie na podstawie danych uzyskanych z innych dokumentów. Jest wiele czynności podczas wypełniania dokumentów, które użytkownicy wykonują ręcznie, podczas gdy dokument na podstawie analizy innych dokumentów może uzyskać automatycznie. Dokumentowi muszą jednak zostać udostępnione dokumenty, z których może pobierać informacje.

3. ARCHITEKTURA DOKUMENTU

Zastosowana architektura dokumentu pozwalająca na tworzenie dokumentów wykonywalnych, mobilnych, interaktywnych i inteligentnych oparta jest na trójwarstwowej architekturze dokumentów IODA [5]. Dokumenty w tej architekturze składają się z następujących warstw:

- danych,
- informacji,
- wiedzy.

Warstwę danych stanowi ciąg bajtów tworzących zawartość plików tekstowych i binarnych, które wchodzą w skład wypełnianego dokumentu. Wzorce, zgodnie z którymi interpretowane są te dane pozwalające na ich interpretacje tworzą warstwę informacji. Natomiast treść, jaka wynika z tych interpretacji stanowi warstwę wiedzy.

4. METODY WYKRYWANIA PLAGIATÓW

Stosowane są powszechnie cztery rodzaje metod wykrywania plagiatów [3]:

1) znajdowanie występujących w różnych dokumentach takich samych n-gramów, gdzie n-gram jest to ciąg kolejnych n znaków występujących w dokumencie [8],
2) analiza stylistyczna i gramatyczna tekstu w celu określenia formy wyrażania treści przez autorów,
3) identyfikowanie zdań różniących się jedynie niektórymi wyrazami, które zostały zastąpione synonimami oraz znajdowanie zdań połączonych lub rozdzielonych,
4) znajdowanie identycznych fragmentów tekstu (ang. fingerprints) występujących w różnych dokumentach.

Ponadto do wykrywania plagiatów stosowane są różne metody określania stopnia podobieństwa między dokumentami oraz fragmentami dokumentów. Używany jest między innymi model przestrzeni wektorowej (ang. Vector Space Model) [9] polegający na tym, że dokumenty reprezentowane są w przestrzeni wielowymiarowej w postaci wektorów. Sposób utworzenia wektora dokumentu zależy od liczby poszczególnych słów, jakie w nim

występują. Podczas obliczania wektorów brane jest również pod uwagę to, czy zawarte w dokumencie słowa występują w innych dokumentach, czy też są rzadkie i dzięki temu charakterystyczne dla dokumentu, w którym wystąpiły.

Dzięki reprezentowaniu dokumentów w postaci wektorów możliwe jest określenie stopnia podobieństwa tych dokumentów. Stopień podobieństwa dwóch dokumentów zależny jest od położenia w przestrzeni określany jest na podstawie podobieństwa między wektorami reprezentującymi te dokumenty. Stosowane są różne miary podobieństwa wektorów. Najpopularniejszą jest miara kosinusowa (ang. cosine similarity), której wartość zależy od kąta między dwoma wektorami, dla których obliczana jest wartość miary.

5. METODA KLASTERYZACJI KIERUNKOWEJ

Metoda klasteryzacji kierunkowej służy do wspomagania użytkowników wyszukiwarek internetowych w wyszukiwaniu informacji [2]. Jest to metoda przeznaczona do interaktywnego rozszerzania zapytań (ang. interactive query expansion). Dzięki niej użytkownicy mogą precyzyjniej określić, co jest przedmiotem ich zainteresowania, dzięki czemu wyszukiwaka ma większe możliwości przedstawienia stron internetowych adekwatnych do oczekiwań użytkownika. Działanie metody klasteryzacji kierunkowej polega na tym, że dla wprowadzonego przez użytkownika zapytania podawane są słowa powiązane tematycznie z tym zapytaniem, które użytkownik może użyć w celu poprawienia swojego zapytania. Słowa te prezentowane są w postaci chmury znaczników (ang. tag cloud). Przykładowo, dla zapytania *car* w języku angielskim wygenerowana zostaje chmura przedstawiona na rysunku 1.

Rys. 1. Chmura znaczników wygenerowana dla zapytania *car*

6. METODA KLASTERYZACJI KIERUNKOWEJ

Dokumenty tworzone w architekturze opracowywanej w projekcie MENAID posiadają funkcjonalność powodującą, że są to dokumenty wykonywalne (patrz rozdz. 3.1). Funkcjonalność dokumentu polegać będzie między innymi na tym, że posiadać on będzie mechanizmy wspierające użytkowników w wypełnianiu dokumentów. Mechanizmy te opierają się na metodach wyszukiwania plagiatów (patrz rozdz. 4) oraz metodzie klasteryzacji kierunkowej (patrz rozdz. 5).

Funkcjonalność wspierająca wypełnianie dokumentu polega na tym, że jeśli użytkownik kliknie prawym przyciskiem myszy w punkcie dokumentu, w którym chciałby uzyskać wsparcie, to na interfejsie dokumentu, w miejscu, gdzie znajduje się wskaźnik myszy, pojawiać się będzie menu zawierające opcję *wyszukaj*. Wskazanie tej opcji spowoduje, że przeprowadzone zostanie wyszukiwanie informacji odnoszących się do wskazanego przez użytkownika miejsca. Informacje te wyszukiwane będą w oparciu o tekst znajdujący się w otoczeniu wskazanego przez użytkownika miejsca. Na przykład, jeśli dokument będzie formularzem, to wyszukiwanie odbywać się będzie na podstawie tytułu pola, które wskazał użytkownik.

Wyszukiwanie przeprowadzane jest w taki sposób, że korzystając ze wskazanych wcześniej metod wyszukiwania plagiatów, znajdowane będą fragmenty innych dokumentów podobne do fragmentu interesującego użytkownika. Ze znalezionych fragmentów wyodrębniane są wyrażenia, które w innych dokumentach znajdują się w miejscu odpowiadającym polu wskazanemu przez użytkownika w wypełnianym przez niego dokumencie. Następnie dokument wykonywalny zaprezentuje te wyrażenia użytkownikowi, który będzie mógł dodać jedno z nich do wypełnianego dokumentu lub na ich postawie sformułować własne wyrażenie.

Dodatkowo, jeśli wskazane w wyszukiwaniu wyrażenia nie będą odpowiadały użytkownikowi, będzie on miał możliwość sprecyzowania rodzaju informacji jakich poszukuje. Przy wynikach wyszukiwania przeprowadzonych metodami identyfikacji plagiatów znajdować się będą również wyniki algorytmu klasteryzacji kierunkowej prezentujące wyrazy powiązane tematycznie z wyszukiwanymi wyrażeniami. Użytkownik będzie mógł wskazać wyrazy, które są adekwatne do jego zainteresowań, dzięki czemu przeprowadzone zostanie ponowne wyszukiwanie wyrażeń w celu zaprezentowania użytkownikowi słów bardziej adekwatnych do jego oczekiwań. Ten proces ponawiania wyszukiwania w oparciu o metodę klasteryzacji kierunkowej użytkownik będzie mógł powtarzać aż do momentu znalezienia wyrażeń, które zdecyduje się wprowadzić do dokumentu.

Przy generowaniu podpowiedzi występuje problem polegający na konieczności określenia, na jakim urządzeniu wykonywany będzie algorytm wyszukujący podpowiedzi na podstawie metod znajdowania plagiatów oraz algorytm klasteryzacji kierunkowej. Jest kilka możliwości. Pierwsza polega na tym, że algorytmy te wykonywane będą na urządzeniu osoby wypełniającej dokument. Druga jest taka, że wykonanie algorytmów odbywać się będzie na zdalnym serwerze. Wyniki tych algorytmów stałyby się wówczas pewną usługą zewnętrzną, z której korzystać mogą użytkownicy wypełniający dokumenty. Trzecia możliwość jest taka, że algorytmy będą częściowo wykonywane na serwerze zdalnym, a częściowo na sprzęcie użytkownika.

Wybór sposobu wykonania algorytmów zależeć będzie od rodzaju urządzenia, z jakiego użytkownik korzysta. W przypadku, gdy do wypełniania dokumentów używane będzie urządzenie przenośne o małej mocy obliczeniowej, algorytmy generujące podpowiedzi uruchamiane będą na serwerze, z którego będą do użytkownika dostarczane jedynie wyniki wyszukiwania.

Uruchamianie w całości wyszukiwania na komputerze należącym do użytkownika, który ma stosunkowo dużą moc obliczeniową też nie jest wskazane z uwagi na to, że wyszukiwanie przeprowadzane jest na znacznej liczbie dokumentów. Konieczne byłoby skopiowanie tych dokumentów na komputer użytkownika w celu dalszego ich przetwarzania. W związku z tym na zdalnym serwerze uruchamiane będą części algorytmów wyszukiwania, które polegają na zidentyfikowaniu dokumentów, w których

potencjalnie znajdować się mogą wyrażenia, którymi może być zainteresowany użytkownik. Dalsze przetwarzanie tych dokumentów odbywać się będzie na urządzeniu użytkownika.

Innym problemem jest określenie zbioru dokumentów, na podstawie którego generowane będą podpowiedzi. Do źródeł tych dokumentów należeć będą wszystkie dokumenty dostępne w Internecie oraz dokumenty zgromadzone w zasobach bibliotek cyfrowych. Jednak dodatkowo ważnym źródłem dokumentów są dokumenty zgromadzone lokalnie na komputerze użytkownika. Często występuje sytuacja taka, że użytkownik wypełnia dokument, który jest podobny do dokumentu, który dawniej wypełniał. Dzięki korzystaniu z dokumentów użytkownika możliwa jest personalizacja podpowiedzi i lepsze dostosowanie ich do potrzeb użytkownika.

7. PODSUMOWANIE

Wprowadzenie architektury dokumentów pozwalającej na tworzenie dokumentów wykonywalnych, mobilnych, interaktywnych i inteligentnych służy przyspieszeniu procesu tworzenia dokumentacji. W celu realizacji tej architektury konieczne jest opracowanie wielu standardów tworzenia takich dokumentów. Między innymi konieczne jest opracowanie standardów dodawania do dokumentu informacji o sposobie jego migracji w Internecie oraz o sposobach zapisu alternatywnych przepływów dokumentu. Istotne jest opracowanie standardów niezależnych od systemu operacyjnego. Powinno być możliwe odczytywanie i wypełnianie dokumentów między innymi na urządzeniach mobilnych. Przedstawiona architektura dokumentu jest przeznaczona dla dowolnego rodzaju dokumentów. Dzięki temu może stać się powszechnie stosowana podczas tworzenia i modyfikowania dokumentów.

8. PODZIĘKOWANIA

Projekt został sfinansowany ze środków Narodowego Centrum Nauki przyznanych na podstawie decyzji numer DEC1-2011/01/B/ST6/06500.

9. BIBLIOGRAFIA

1. Strona internetowa projektu MENAID (MEtody i NArzędzia Inżynierii Dokumentu przyszłości) http://menaid.org.pl/index.php/pl/
2. Kaczmarek A. L.: Interactive Query Expansion With the Use of Clustering-by-Directions Algorithm, IEEE Transactions on Industrial Electronics, Vol. 58, No. 8, IEEE August 2011, s. 3168 – 3173, ISSN: 0278-0046.
3. Pera M. S., Ng Y.-K.: SimPaD: A word-similarity sentence-based plagiarism detection tool on Web documents, Web Intelligence and Agent Systems, Vol. 9 Issue 1, Amsterdam IOS Press January 2011, s. 27-41, ISSN 1570-1263.
4. Executable Paper Grand Challenge, Elsevier 2011, http://www.executablepapers.com/about-challenge.html
5. Siciarek J., Wiszniewski B.: IODA - an Interactive Open Document Architecture, Procedia Computer Science, Proceedings of the International Conference on Computational Science, ICCS 2011, Vol. 4, Elsevier 2011 s. 668-677, ISSN 1877-0509
6. Godlewska M. Wiszniewski B.: Distributed MIND – A New Processing Model Based on Mobile Interactive Documents, Parallel Processing and Applied Mathematics, LNCS Vol. 6068, Springer Berlin / Heidelberg 2010 s. 244-249, ISBN 978-3-642-14402-8
7. Godlewska M.: Agent System for Managing Distributed Mobile Interactive Documents, Agent and Multi-Agent Systems: Technologies and Applications, LNCS Vol. 6071, Springer Berlin/Heidelberg 2010 s. 390-399, ISBN 978-3-642-13540-8
8. Cavnar W. B., Trenkle J. M..: N-gram-based text categorization, Proceedings of SDAIR 94, 3rd Annual Symposium on Document Analysis and Information Retrieval, USA Las Vegas Nevada 1994, s. 161–175.
9. Salton G.: Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer, Addison-Wesley Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA, 1989, ISBN 978-0201122275.

METHODS OF PROVIDING INFORMATION FOR MOBILE INTERACTIVE DOCUMENTS BASED ON DETECTING SIMILARITIES IN CONTENTS OF DOCUMENTS

Key-words: electronic documents, information retrieval

The paper is concerned with the novel architecture of the document which provides more efficient techniques of document processing. Documents prepared with the use of this architecture are executable, mobile, interactive and intelligent. They cooperate with users in the process of gathering information and creating content of the document. Moreover, documents are able to automatically migrate in the Internet to users who have the required information. These documents contain not only text, pictures and other information in a graphical form. They also have functionality which can execute contents available in documents. This paper presents the usage of methods for plagiarism detection and information retrieval methods to facilitate users in creating these kinds of documents. In particular, the paper presents the application of clustering by direction method designed to support users of search engines.