TECHNIKA TRANSPORTU SZYNOWEGO

Urszula KUŻELEWSKA, Wojciech RAKOWSKI

INTELIGENTNA WYSZUKIWARKA INTERNETOWA OPARTA NA GRUPOWANIU DOKUMENTÓW

Streszczenie

Powstanie sieci WWW spowodowało w ostatnich latach wzrost dostępności specjalistycznych informacji dla każdego użytkownika komputera podłączonego do Internetu. Liczba dostępnych w Internecie informacji rośnie w ogromnym tempie. Wraz ze wzrostem liczby stron internetowych trudniejszym staje się odnalezienie poszukiwanej informacji. Odpowiedzią na ten problem było powstanie wyszukiwarek internetowych, które na zapytanie użytkownika opisujące poszukiwaną informacje zwracają listę dokumentów mniej lub bardziej odpowiadających zapytaniu. Wyszukiwarki internetowe nie są idealnym rozwiązaniem, ponieważ zwrócona lista dokumentów jest długa i często zawiera dokumenty nie związane z poszukiwaną informacją. Grupowanie dokumentów jest rozwiązaniem mającym na celu poprawę jakości prezentacji wyników wyszukiwania, gdyż umożliwia wyświetlenie ich w postaci tematycznie powiązanych grup. W artykule przedstawiono wyniki grupowania dokumentów z sieci WWW zwrócone przez jedną z popularnych wyszukiwarek. Wykorzystano następujące metody grupujące: EM i AHC.

WSTĘP

Internet jest popularnym miejscem do dzielenia się naszą wiedzą lub wyrażania własnej opinii, jak również źródłem ciekawych i wartościowych informacji. Pomimo istnienia niesamowicie dużej liczby stron internetowych, staną się one z czasem bezużyteczne, jeśli będą trudne do znalezienia. "Rewolucja Internetu w dziedzinie dostępu do informacji nie polega na ich większej ilości (ogromne ilości informacji od dawna były dostępne w bibliotekach i innych miejscach), ale na zwiększeniu efektywności w dostępie do nich" [5, str. 107-109]. Jednakże większa ilość informacji nie oznacza wzrostu jej jakości. Ten problem stanowi wyzwanie dla współczesnych wyszukiwarek, które muszą się nieustannie rozwijać, aby oferować wysoki standard generowanych wyników.

Jednym z rozwiązań tego problemu jest wykorzystanie metod SRC (Search Results Clustering) służących do identyfikacji grup podobnych stron internetowych. Według Weissa [10] wynikiem działania takich algorytmów są grupy dokumentów, które są mogą mieć wyodrębniony wspólny tematu i są opisane zrozumiale dla człowieka. Takie podejście nie wpływa na jakość i długość listy wyników, jednakże zmniejsza czas dostępu do istotnych informacji.

Celem niniejszego artykułu jest przedstawienie możliwości wykorzystania grupowania w dziedzinie wyszukiwania w sieci Web. Sama idea stosowania algorytmów grupujących

do organizacji dokumentów nie jest nowa, jednak wciąż proponuje się wiele nowych metod zarówno w dziedzinie analizy skupień, jak przetwarzania dokumentów.

Artykuł jest zorganizowany w następujący sposób: pierwszy rozdział opis dziedziny przetwarzania dokumentów tekstowych, drugi - krótki przegląd metod grupowania, natomiast kolejny zawiera wyniki eksperymentów. Ostatni rozdział podsumowuje artykuł.

1. PRZETWARZANIE DOKUMENTÓW

Wynikiem wyszukiwania zwróconym przez wyszukiwarkę internetową jest lista dokumentów w sieci WWW podobna do wprowadzonego przez użytkownika dokumentu, którym jest również zapytanie. Każdy element z listy zawiera tytuł strony, jej adres oraz krótki tekst pobrany z oryginalnej treści (tzw. snippet). Wstępne przetwarzanie tekstu (przede wszystkim ze snippetów, ale również z tytułów) składa się z następujących etapów: ujednolicenie wielkości liter, usunięcie słów z tzw. stop listy, stemming, selekcja termów oraz sposoby reprezentacji dokumentu w przestrzeni wektorowej. Wstępne przetwarzanie korzystnie wpływa na czas całego procesu oraz zwiększa stopień podobieństwa pomiędzy dokumentami.

Stop lista składa się ze słów, które nie zawierają znaczącej treści i nie są związane z tematem dokumentu. Przykładami takich słów są: „jest", „on", „wykonać" (język polski) lub „is", „the", „but" (język angielski). Lista takich słów jest przydatna w przypadku, gdy należy zidentyfikować język dokumentu, ale w dalszym procesie przetwarzania są zbędne i należy je usunąć.

Stemming polega na ujednoliceniu słów o podobnej treści, lecz w różnej formie językowej. W wyniku tego są generowane krótsze, lecz jednakowe dla różnych form słów, wyrazy zwane termami. Przykładowo: „liczyć", „liczydło" i „liczba" mogą być zastąpione termem „licz".

Dokument opisany termami może być przetransformowany do przestrzeni wektorowej (VSM – Vector Space Model) [7, str. 613-620], w której jest on reprezentowany *n* liczbami (patrz Równanie 1). Atrybutami każdej strony internetowej w VSM jest zbiór *n* termów ze wszystkich dokumentów. W ten sposób można do grupowania tekstu wykorzystywać metody operujące danymi numerycznymi. Liczby, które zastępują termy są związane z istotnością pierwotnie obecnych tam słów.

$$D_i = (d_{i1}, d_{i2}, \ldots, d_{in}) \tag{1}$$

Istotność słów w dokumentach wyraża się na wiele sposobów. Jednym z nich jest reprezentacja binarna, w której wartością „1" oznacza się wystąpienie danego słowa w dokumencie. W przeciwnym przypadku wstawia się „0". Inne rozwiązania opierają się na częstości występowania poszczególnych termów w dokumentach (TF – Term Frequency) lub zbiorze dokumentów (DF - Document Frequency). Często wykorzystywanym podejściem jest jednoczesne uwzględnienie częstości termu w dokumencie oraz w całym zbiorze danych (TFIDF - Term Frequency Inverse Document Frequency).

Metody selekcji podzbioru termów opierają się głównie o wartości atrybutów w przestrzeni VSM. Można wybrać jedynie te słowa, które występują najczęściej w dokumencie lub w całym zbiorze, można też sugerować się wysoką wartością współczynnika TFIDF.

Etapem, który realizuje się po grupowaniu jest generowanie etykiet grup. Do tej pory zaproponowano wiele takich metod. Najprostsze z nich bazują na pobraniu do etykiety najczęściej powtarzających się słów w danej grupie. Bardziej zaawansowane mogą w tym celu analizować ciągi kilkuwyrazowe zwane n-gramami.

2. GRUPOWANIE DOKUMENTÓW

Celem grupowania jest identyfikacja grup podobnych do siebie elementów z danych wejściowych. Analiza podobieństwa odbywa się na podstawie wewnętrznych zależności i powiązań, bez obecności atrybutu przynależności do grupy [3, str. 264–323]. Tradycyjne metody grupowania wykorzystano m.in. w następujących rozwiązaniach: WISE [1, str. 301-304], HKM [6, str. 370–391].

Metody grupujące dzieli się na techniki partycjonujące, hierarchiczne, oparte na modelu oraz na gęstości [2]. Pierwsze z nich działają na zasadzie przydziału każdego z elementów do jednej z grup i sprawdzaniu efektywności takiego rozwiązania (przykładowo: k-średnich, k-modów). Metody hierarchiczne tworzą dendrogram, czyli hierarchiczną strukturę odwzorowującą zależności pomiędzy danymi (np. AHC, DHC). Algorytmy oparte na modelu zakładają wybrany model danych i na podstawie aktualnego zbioru danych oszacowują parametry tego modelu (np. EM, SOM). Metody oparte na gęstości jako grupy traktują skupiska o większym zagęszczeniu obiektów, które są przez nie identyfikowane. Przykładem takiej techniki jest DBSCAN.

1.1. Algorytm EM

W algorytmie EM (z ang. Expectation Maximization) założono model danych złożony z grup tworzących mieszaninę rozkładów normalnych [4, str. 191–201]. Dla tego modelu estymowane są optymalne, najbardziej prawdopodobne wartości parametrów, takie jak wartość średnia i macierz kowariancji. Algorytm składa się dwóch kroków: estymacji i maksymalizacji, które są wykonywane iteracyjnie. W pierwszym z nich wyznacza się prawdopodobieństwa przynależności wszystkich obiektów do poszczególnych grup dla aktualnych wartości parametrów rozkładów, a następnie wyznacza się ich nowe najbardziej prawdopodobne wartości.

Wśród użytecznych właściwości metody znajduje się umiejętność przetwarzania danych z brakującymi wartościami atrybutów, ponieważ w metodzie wykorzystano algorytm przetwarzania wiedzy niepewnej według teorii Dempstera. Dodatkowym jej atutem jest zdolność detekcji liczby istniejących w zbiorze skupień.

1.2. Algorytm AHC

Hierarchiczne techniki aglomeracyjne (AHC) zaczynają działanie od analizy zależności pomiędzy pojedynczymi obiektami, idąc w górę drzewa łączą je w coraz większe grupy, natomiast techniki dzielące (DHC) zaczynają od całego zbioru połączonego w jedną grupę, dzieląc go na mniejsze zbiory na kolejnych niższych poziomach hierarchii.

W metodach hierarchicznych można zastosować różne sposoby wyliczania odległości pomiędzy grupami. Są to m.in.: algorytm pojedynczego połączenia (z ang. single-link), zupełnego połączenia (z ang. complete-link) oraz minimalnej wariancji (z ang. minimum-variance) [2].

Algorytmy hierarchiczne są popularnymi technikami grupującymi pozbawionymi szeregu wad obecnych w podejściu partycjonującym, jednakże ich złożoność obliczeniowa jest bardzo duża. Oprócz tego, wykazują one dodatkową niedogodność. Proces łączenia w grupy (podziału – w algorytmach dzielących) jest ostateczny na poszczególnych poziomach, więc w kolejnych iteracjach nie jest możliwa poprawa, ani dostrojenie już istniejącego wyniku. Wymaga się również podawania liczby grup do podziału.

Ze względu na szereg opcjonalnych rozwiązań szacowania odmienności pomiędzy grupami umożliwiających dostosowanie do różnych typów zbiorów, podejście hierarchiczne jest bardzo często wykorzystywane w eksploracji danych.

2. WYNIKI BADAŃ

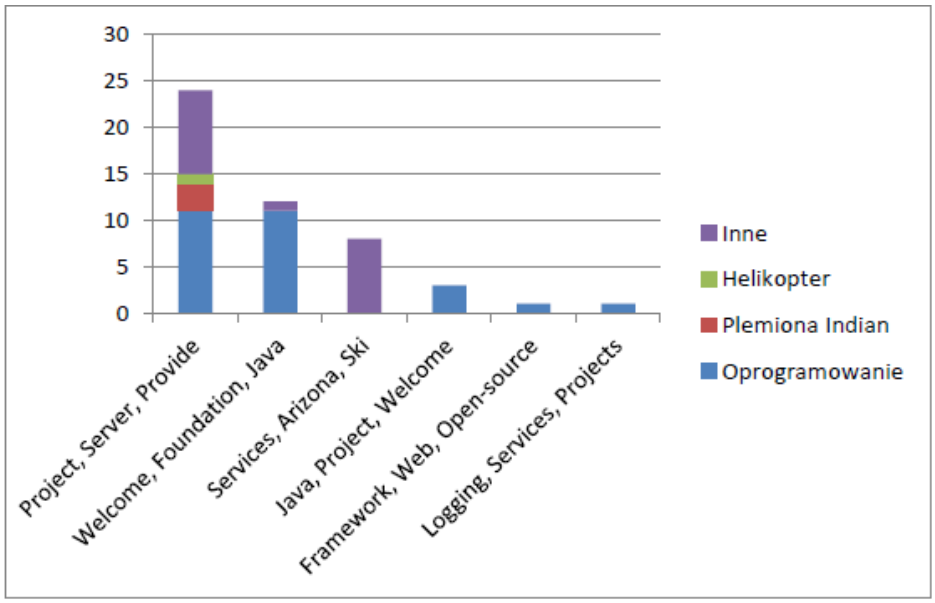
Oprogramowanie użyte do poniższych eksperymentów powstało w ramach pracy magisterskiej [8] powstałej w Politechnice Białostockiej. Wyniki wyszukiwania pozyskano z wyszukiwarki Bing [12], natomiast algorytmy grupujące pochodzą z systemu Weka [11]. W programie istnieje również możliwość grupowania danych z pliku.

W niniejszym artykule zamieszczono wyniki dwóch eksperymentów: grupowania wyników uzyskanych z zapytania „apache" wprowadzonych do wyszukiwarki Bing oraz grupowania wyników kilku zapytań zamieszczonych w pliku. Wykorzystano dwie metody grupowania EM oraz AHC. EM jest metodą działającą z zadaną wartością liczby grup do wydzielenia, jednakże w systemie Weka zaimplementowano możliwość automatycznej jej detekcji.

2.1. Grupowanie wyników zapytania z wyszukiwarki

W tym eksperymencie do wyszukiwarki wprowadzono zapytanie „apache" i otrzymano listę wyników zawierającą m. in. dokumenty dotyczące oprogramowania, plemion Indian oraz helikoptera. Algorytmom ustawiono następujące parametry: metoda opisu w przestrzeni dokumentów – TFIDF, długość wektora atrybutów – 30 termów, niezdefiniowana liczba grup (EM) lub liczba grup równa 8 (AHC).

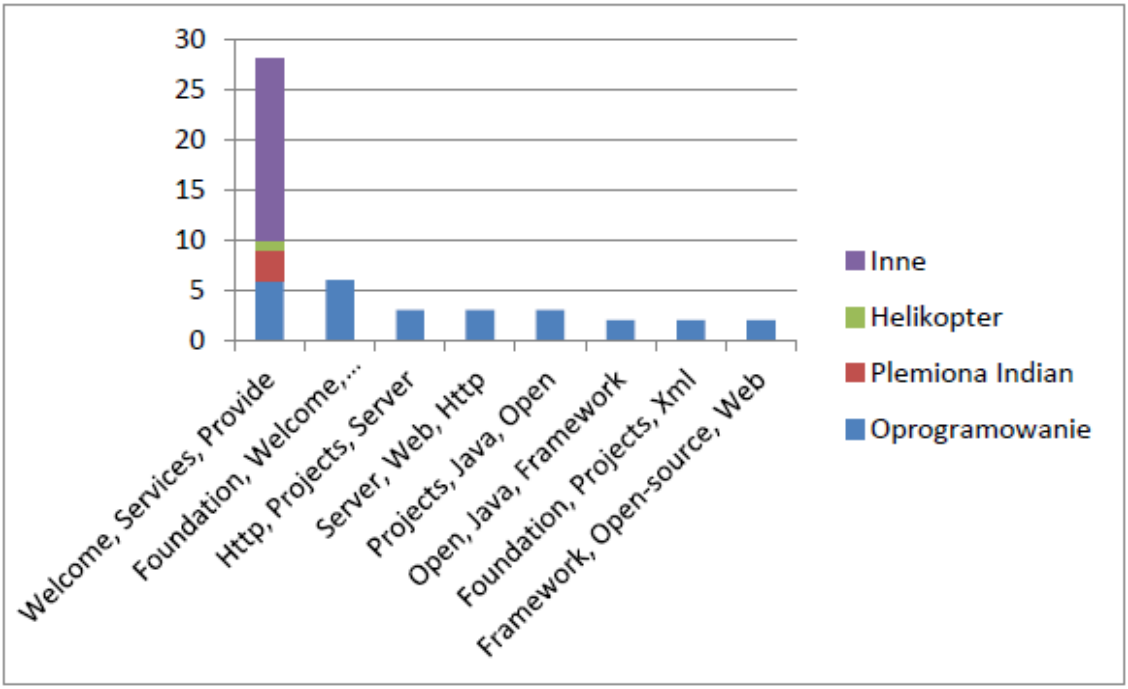
Rysunki 1 oraz 2 przedstawiają wyniki wyszukiwania otrzymane z wyszukiwarki Bing dla zapytania „apache" wygenerowane przez odpowiednio algorytmem EM oraz AHC (liczności grup oraz ich etykiety tematyczne). Odpowiednimi kolorami oznaczono dokumenty z w.w. tematów.



Rys. 1. Podział na grupy wyników wyszukiwania otrzymanych z zapytania „apache" otrzymany algorytmem EM

Podział otrzymany z wykorzystaniem algorytmu EM charakteryzuje się dobrą separacją stron powiązanych z oprogramowaniem, jednakże strony dotyczące Indian i helikoptera znalazły się w jednej grupie razem z częścią wyników dotyczących firmy Apache. Powodem

tego jest przeważająca liczność w liście wyników dokumentów związanych z systemami informatycznymi (55%).



Rys. 2. Podział na grupy wyników wyszukiwania otrzymanych z zapytania „apache" otrzymany algorytmem AHC

Podział otrzymany z wykorzystaniem algorytmu AHC również zawiera grupy związane jedynie z oprogramowaniem, lecz liczność ich jest nieduża w porównaniu z licznością największej grupy, w której znalazły się dokumenty dotyczące helikoptera i Indian.

2.2. Grupowanie wyników wielu zapytań z pliku

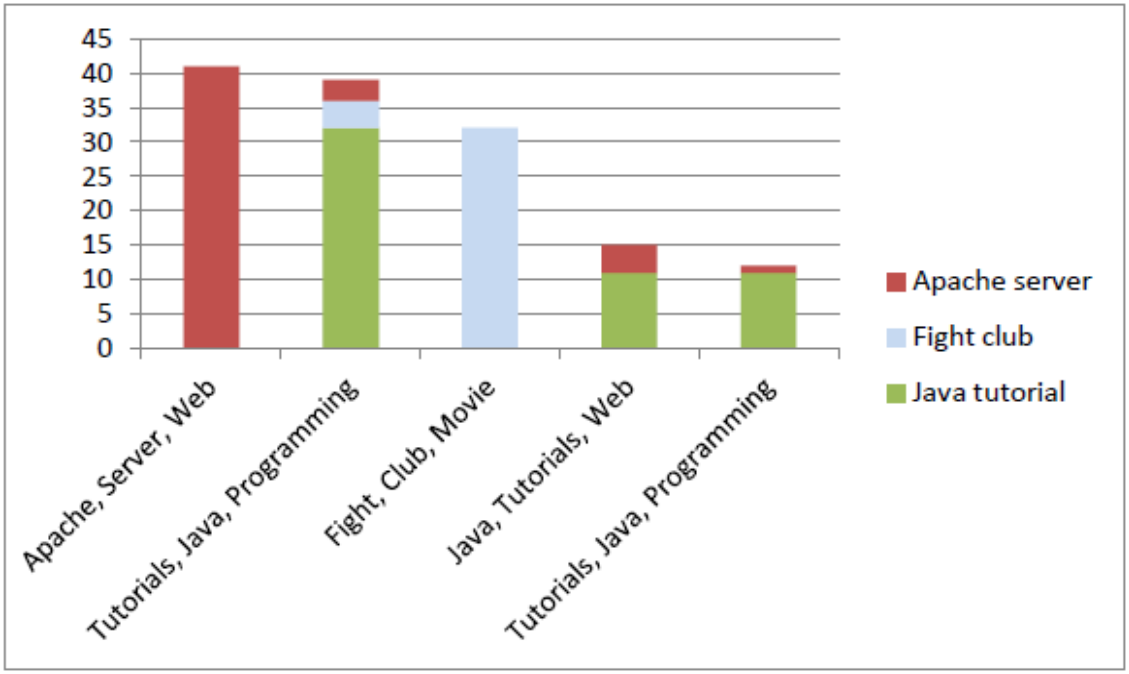
Dane do tego eksperymentu pozyskano następująco: do wyszukiwarki Bing wprowadzono kilka zapytań i każdorazowo pobierano zwrócone wyniki o określonej liczbie, zapisując do pliku te strony, które były ściśle powiązane z zapytaniem. Tabela 1 zawiera zestawienie treści zapytań oraz liczby dokumentów zapisanych do pliku.

Tab. 1. Treści zapytań oraz liczbą dokumentów związanych z zapytaniem pobrana z wyników wyszukiwarki

Zapytanie	Opis wyszukiwanej informacji	Liczba dokumentów
Fight club	Film pt. „Fight Club" w reżyserii Davida Finchera	36
Apache server	Serwer stron WWW firmy Apache	49
Java tutorial	Zasady tworzenia i przykłady kodu w języku obiektowym Java	54

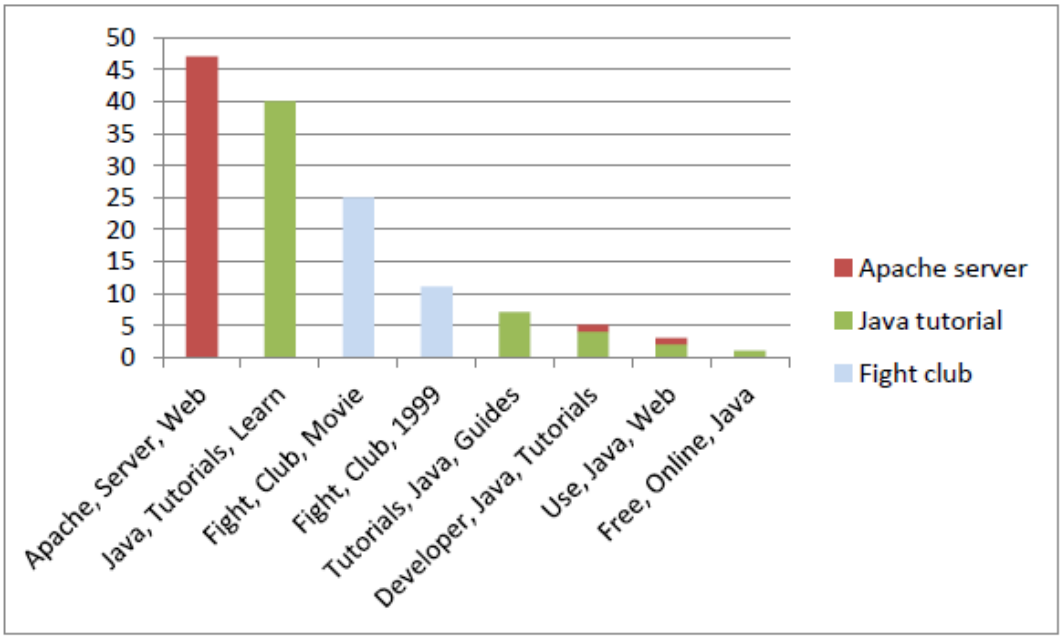
Pierwsze z zapytań dotyczy filmu Davida Finchera „Fight club", drugie – firmy produkującej systemy informatyczne Apache, natomiast trzecie języka oprogramowania Java. Dwa ostatnie zapytania dotyczą zbliżonej tematyki, co jest kolejną trudnością dla metody grupującej. Algorytmom ustawiono parametry z poprzedniego eksperymentu.

Rysunki 3 i 4 przedstawiają podział na grupy wyników wyszukiwania otrzymanych z wielu zapytań wygenerowany odpowiednio algorytmem EM i AHC.



Rys. 3. Podział na grupy wyników wyszukiwania otrzymanych z wielu zapytań wygenerowany algorytmem EM

Algorytm EM wygenerował 5 grup, z których dwie były homogeniczne: grupa dotycząca serwera Apache (o etykiecie „Apache, Server, Web") oraz filmu („Fight, Club, Movie"). Pozostałe w znacznej większości dotyczyły języka Java (dwie o etykietach „Tutorials, Java, Programming" i „Java, Tutorials, Web"). Etykiety grup dotyczących Javy są zbyt podobne i powinny tworzyć jedną grupę.



Rys. 4. Podział na grupy wyników wyszukiwania otrzymanych z wielu zapytań wygenerowany algorytmem AHC

Podział wygenerowany przez metodę AHC zawierał 8 grup, z których 6 było homogenicznych („Apache, Server, Web", „Java, Tutorial, Learn", „Fight, Club, Movie", „Fight, Club, 1999", „Tutorials, Java, Guides" oraz „Free, Online, Java"), a dwie w większości dotyczyły stron o języku Java („Developer, Java, Tutorials" oraz „Use, Java, Web"). W tym przypadku wyniki, choć zawierały większą liczbę grup, były lepsze (większa liczba grup homogenicznych oraz bardziej zróżnicowane zawartości i etykiety) niż w przypadku algorytmu EM.

Otrzymane wyniki są poprawne. Każdy z algorytmów utworzył odmienny podział, jednakże wydzielone grupy są w większości związane z jednym tematem.

3. PODSUMOWANIE

Celem niniejszego artykułu było przedstawienie możliwości zastosowania metod grupowania w dziedzinie wyszukiwania w sieci Web. Do eksperymentów wybrano dwa algorytmy: EM i AHC. Otrzymane wyniki oceniono pod względem homogeniczności ich zawartości.

Niewątpliwie można wprowadzić wiele udoskonaleń w prezentowanej metodzie, jednakże zamieszczone wyniki wskazują na dużą efektywność tradycyjnych metod grupowania w dziedzinie SRC. W doświadczeniu pierwszym, oprócz grupy zawierającej dokumenty o różnych tematach, znalazły się również skupienia związane jedynie z oprogramowaniem. W eksperymencie drugim skuteczność obu metod była znacznie wyższa, gdyż liczność wyników zapytań była zbliżona. Utworzone grupy były w większości homogeniczne.

Z powyższych eksperymentów można wyciągnąć wniosek o dużych możliwo-ściach metod grupujących zastosowanych do identyfikacji dokumentów podobnych. Warte sprawdzenia są wszystkie: istniejące oraz nowo proponowane algorytmy, m.in. ze względu na ciągłe udoskonalenia w dziedzinie przetwarzania dokumentów.

AN INTELLIGENT WEBSEARCHING BASED ON DOCUMENT CLUSTERING

Summary

Development of the World Wide Web over recent years led to increased availability of specialized information for each user with a computer connected to the Internet. The amount of information available there is increasing rapidly and finding desirable information is more difficult. The solution of the problem may be Internet search engines, however they have some disadvantages. They require from users to input a query describing searching information and they return a list of documents, which is very long and often contains websites not relevant to the query. To increase efficiency of the searching process one may identify groups of similar documents from a result list. One of the tools to do it are clustering algorithms. The article presents clustering of Web search results from one of the popular search engines grouped using the following methods: EM and AHC.

BIBLIOGRAFIA

1. Campos, R., Dias, G., Nunes, C.: *WISE: Hierarchical Soft Clustering of Web Page Search Results based on Web Content Mining Techniques*, Konferencja IEEE/WIC/ACM International Conference on Web Intelligence, 2006
2. Jain, A.K., Dubes, R.C.: *Algorithms for clustering data*, Prentice Hall, 1988

3. Jain, A.K., Murty, M.N. i Flynn, P.J.: *Data clustering: a review*, ACM Computing Surveys 1999, nr 31(3)
4. Lauritzen, S.L.: *The EM algorithm for graphical association models with missing data*, Computational Statistics and Data Analysis 1995, nr 19
5. Lawrence, S., Giles, C. L.: *Accessibility and Distribution of Information on the Web*, Nature 1999, nr 400
6. Mahdavi, M., Abolhassani, H.: *Harmony K-means algorithm for document clustering*, Data Mining and Knowledge Discovery 2009, nr 18
7. Markow, Z., Larose, D.: *Eksploracja zasobów internetowych*, PWN, 2009
8. Rakowski, W.: Inteligentna wyszukiwarka internetowa wykorzystująca grupowanie w celu optymalizacji wyników wyszukiwania, praca magisterska, Politechnika Białostocka, 2011
9. Salton, G.: *A Vector Space Model for Automatic Indexing*, Communications of the ACM, 1975, nr 18(11)
10. Weiss, D.: *O szukaniu igły w stogu siana*, Seminarium Instytutu Lingwistyki, Polska Akademia Nauk, 2003
11. System Weka, http://www.cs.waikato.ac.nz/ml/weka/, (dostęp 26.09.2012)
12. Wyszukiwarka Bing, http://www.bing.com, (dostęp 10.10.2012)

Autorzy:
dr inż. Urszula KUŻELEWSKA– Politechnika Białostocka
mgr inż. Wojciech RAKOWSKI– absolwent Politechniki Białostockiej

Artykuł zrealizowano w ramach pracy badawczej nr S/WI/5/08.