JULIAN SZYMAŃSKI, WŁODZISŁAW DUCH

Wydział Elektroniki Telekomunikacji i Informatyki, Politechnika Gdańska
julian.szymanski@eti.pg.gda.pl
Wydział Fizyki, Astronomii i Informatyki Stosowanej, Uniwersytet Mikołaja Kopernika w Toruniu
wduch@is.umk.pl

Wizualizacja struktury Wikipedii do wspomagania wyszukiwania informacji

1. Wstęp

Graficzna prezentacja jest efektywnym sposobem poprawiania interakcji użytkownika z repozytorium wiedzy. Pozwala ona na przejrzyste przedstawienie złożonych struktur i uchwycenie zależności, które nie są widoczne bezpośrednio. Zastosowanie takiego podejścia w wyszukiwaniu informacji pozwala na prezentację danych na wysokim poziomie abstrakcji przy jednoczesnym określeniu ich kontekstu, co ma bezpośrednie przełożenie na jakość dostępu do informacji.

Istniejące aktualnie wyszukiwarki internetowe wyświetlają rezultaty swojej pracy w formie obszernej listy wyników. Sposób ten jest często wystarczający do odnalezienia specyficznych treści, jednak wykazuje trudności przy zapytaniach wieloznacznych. Nie umożliwia on określenia dziedziny wiedzy, która ma być przeszukana, nie można też zastosować wyszukiwania opartego na podobieństwie do zadanych przykładów. W dużych repozytoriach wiedzy przydatna jest również graficzna wizualizacja wyników wyszukiwania wraz z przejrzystym podziałem na kategorie.

2. Wikipedia jako repozytorium wiedzy

Wikipedia jest obecnie jednym z największych zbiorów wiedzy ludzkiej. Ze względu na swój rozmiar typowe dla encyklopedii wyszukiwanie alfabetyczne lub korzystanie z indeksu nie jest tu wystarczające. Dostęp do informacji zawartych w tak dużej encyklopedii wymaga użycia bardziej efektywnych mechanizmów. Najprostszym rozszerzeniem oferowanym przez Wikipedię jest wyszukiwanie po słowach kluczowych wynajdywanych zarówno w tytule artykułu, jak i w jego treści. Ze względu na wieloznaczności

występujące w języku naturalnym podejście to również nie wystarcza. Dla nagłówków artykułów kompensowane jest to stronami ujednoznaczniającymi, jakiej dziedziny dotyczy dany artykuł, co rozszerza kontekst poszukiwania. Częstym problemem w wyszukiwaniu informacji jest również sytuacja, kiedy szukający nie potrafi dokładnie określić, czego poszukuje, a jedynie z grubsza możliwe jest określenie interesującego go obszaru. Innym istotnym zagadnieniem jest wyszukiwanie informacji podobnej do znanych treści.

By wspomóc rozwiązanie tych problemów, w Wikipedii budowany jest system kategorii umożliwiający przeglądanie repozytorium z użyciem abstrakcyjnych pojęć, które łączą artykuły w powiązane tematycznie grupy. W celu poprawy przeglądania repozytorium Wikipedii i wyszukiwania w nim informacji zaproponowaliśmy odpowiednie narzędzia.

3. Narzędzia do wizualizacji

3.1 Narzędzie do prezentowania struktury kategorii

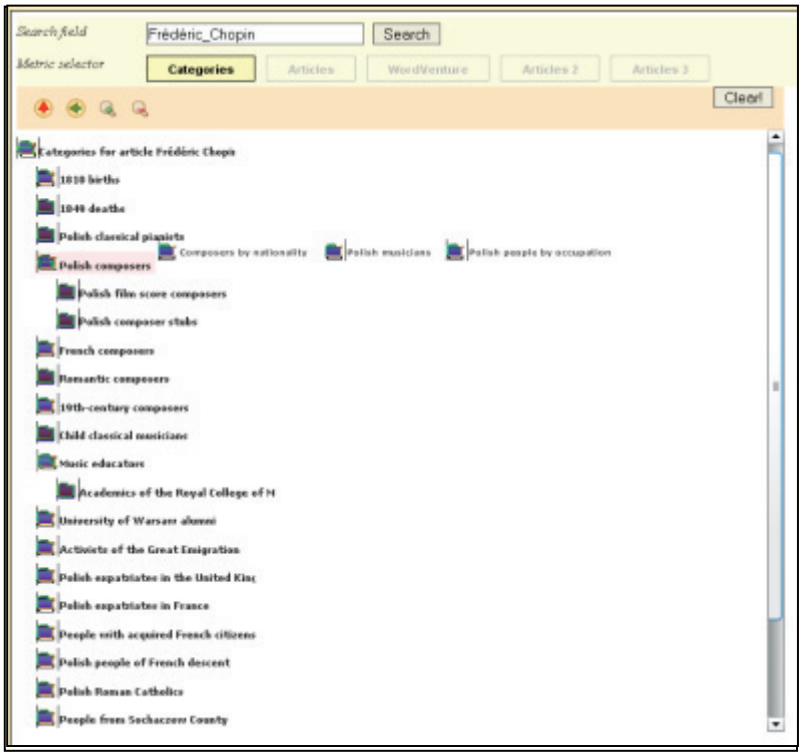
Kategorie Wikipedii tworzą strukturę podobną do hierarchicznej. Ze względu na fakt, że nie jest to idealne drzewo: nie występują w nim cykle i każdy z węzłów podrzędnych ma tylko jeden węzeł nadrzędny, nie można systemu kategorii przeglądać jak prostego katalogu. Konieczne jest również rozwiązanie sytuacji, w których podkategorie mogą należeć do wielu nad-kategorii. Na rysunku 1. przedstawiono graficzną prezentację struktury kategorii opisujących artykuł „Fryderyk Chopin". Nasz autorski komponent umożliwia prezentacje kategorii, do których artykuł przynależy bezpośrednio, jak również, poprzez powiązania z innymi kategoriami, pozwala poruszać się pomiędzy nimi i przeglądać artykuły w nich zawarte.

Komponent osadzony jest jako aplikacja internetowa, która pośredniczy w komunikacji pomiędzy użytkownikiem a serwerami Wikipedii. Stanowi swego rodzaju nakładkę transformującą już istniejące w Wikipedii informacje i prezentującą je użytkownikowi w bardziej przyjazny sposób. Umożliwia poruszanie się po kategoriach poprzez graficzny interfejs tak, że przypomina to nawigowanie po systemie plików i katalogów.

3.2 Graficzna prezentacja podobieństwa artykułów

Wyszukiwanie informacji zbliżonej do zadanej odbywa się w oparciu o określoną miarę podobieństwa. Miara ta może być

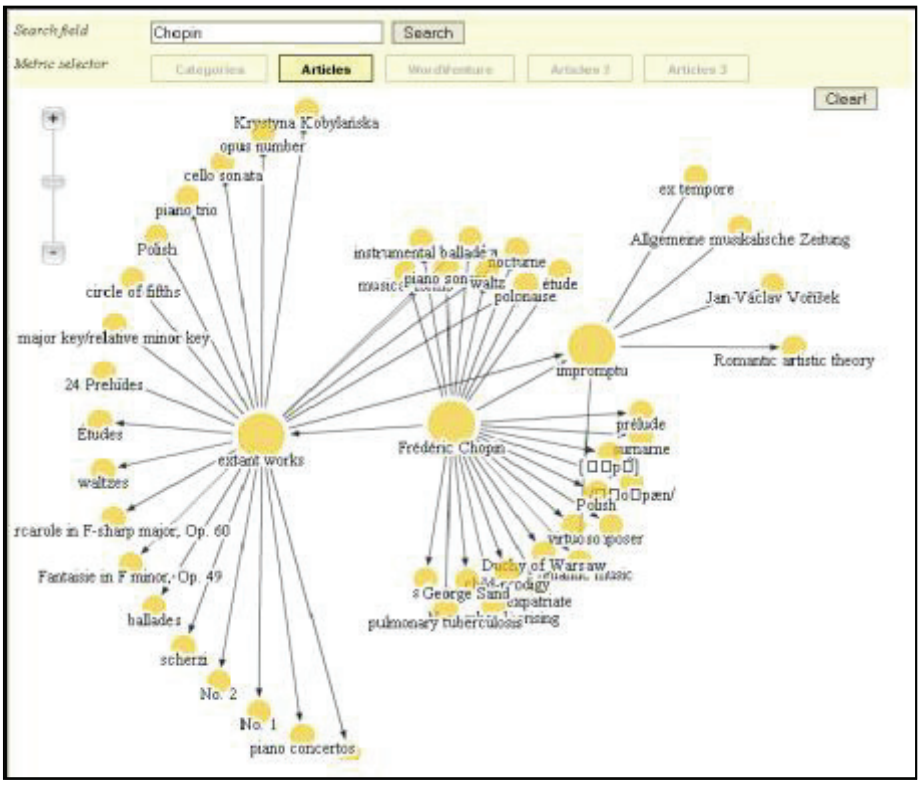
zbudowana na bardzo wiele sposobów, opisując różne aspekty podobieństwa, jakie mogą zachodzić między dokumentami tekstowymi.



Rys. 1. Komponent do przeglądania kategorii Wikipedii.

Do utworzenia jednej z najprostszych miar podobieństwa dla artykułów Wikipedii można użyć linków występujących pomiędzy artykułami. W podejściu tym najbardziej podobnymi do zadanego artykułu są te artykuły, które posiadają referencje w postaci hiperpowiązania. Podobieństwo oparte na powiązaniach może zostać zobrazowane w postaci grafu. Do przedstawienia artykułów Wikipedii wykorzystaliśmy drugi nasz komponent umożliwiający poruszanie się po dużych grafach w interaktywny sposób. Interaktywność realizowana jest tu przez umożliwienie użytkownikowi wskazania węzła, który jest dla niego interesujący, oraz wyświetlenie wokół niego innych, najbardziej podobnych. Miara podobieństwa określona jest tu krawędzią grafu, którego wierzchołki oznaczają artykuły. Ze względu na rozmiar grafu wyświetlany jest tylko jego niewielki fragment. Wskazanie węzła powoduje ustalenie go jako centralnego, a następnie wyświetlenie elementów skojarzonych z nim, które poprzez podobieństwo

określają kontekst. Na rysunku 2. przedstawiono fragment grafu Wikipedii zbudowanego poprzez wzajemne powiązania artykułów oparte na referencjach. W przykładzie centralnym węzłem jest artykuł o Fryderyku Chopinie, natomiast w jego otoczeniu przedstawione zostały artykuły, do których ma on odnośniki.

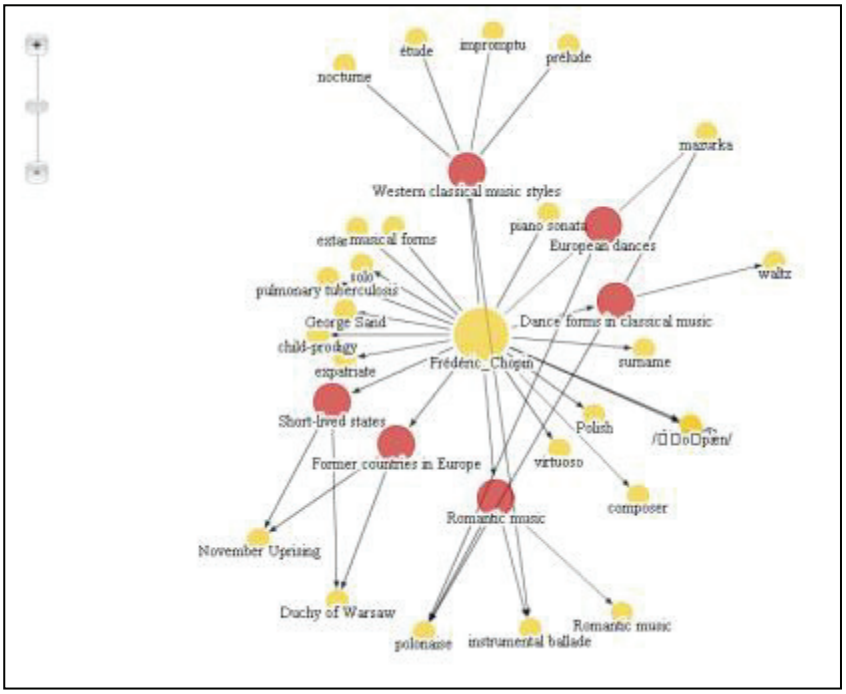


Rys. 2. Wizualizacja struktury powiązań pomiędzy artykułami z użyciem interaktywnego grafu.

W przedstawionej tu wizualizacji referencje między artykułami określają ich podobieństwa. Taką miarę można zbudować w prosty sposób przez wydobycie z treści strony artykułu znaczników *href*. W przyszłości planujemy rozwinięcie tej metody, konstruując inne miary wykorzystujące własności języka naturalnego, tak by podobieństwo pomiędzy artykułami mogło być określone w sposób semantyczny. Realizacja tego w najprostszej formie odbywać się może poprzez dodanie do krawędzi w grafie typu określającego powiązanie między dwoma węzłami. Taka graficzna prezentacja umożliwia wykorzystanie różnych metryk podobieństwa, co pozwala na przedstawienie artykułów skojarzonych pod różnym kątem.

3.3 Wizualizacja powiązań z użyciem kategorii

W sytuacji gdy artykuł ma wiele referencji (lub jest wiele dokumentów powiązanych z nim inną metryką podobieństwa), wyświetlenie ich wszystkich może spowodować, że wizualizacja będzie mało czytelna. Podejście proponowane w paragrafie 3.2 rozszerzone zostało o powiązanie z systemem kategorii. W wizualizacji przedstawionej na rysunku 3. zaprezentowano jedynie artykuły podobne do zadanego, które mają więcej niż jedno wzajemne odniesienie. Oznaczone są one krawędziami wychodzącymi bezpośrednio z artykułu centralnego. Dodatkowo na wizualizacji tej przedstawiono kolorem czerwonym kategorie, które wiążą inne artykuły. Przyjęto, że wyświetlone zostają tylko kategorie z artykułami, do których występują odnośniki z artykułu centralnego. W celu ograniczenia liczby kategorii i odrzucenia artykułów mniej istotnych wyświetlane są tylko te, które mają więcej niż jeden artykuł, do którego jest referencja z artykułu centralnego.

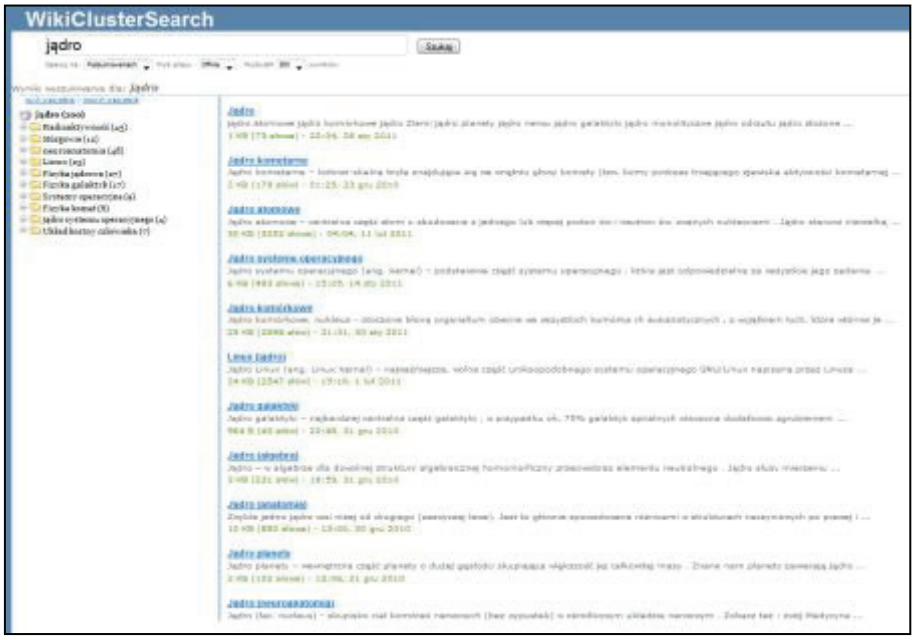


Rys. 3. Prezentacja graficzna powiązań między artykułami wykorzystująca informacje kategorialne.

3.4 Wyszukiwanie przez grupowanie

Przedstawione powyżej sposoby wizualizacji mogą być użyteczne w sytuacji, gdy użytkownik do interakcji z repozytorium wiedzy wykorzystuje jedynie wskazywanie myszą. Wizualizacja taka pozwala na swobodne poruszanie się po repozytorium i przeglądanie artykułów skojarzonych z zadanym wedle określonej miary podobieństwa.

W ramach naszych prac badawczych rozwijamy również metody dostępu do informacji rozszerzające typowe wyszukiwanie oparte na słowach kluczowych. W tradycyjnym podejściu użytkownik zadaje słowa kluczowe, które zostają wyszukane w repozytorium, a jako rezultat zwrócona zostaje lista artykułów zawierających zadane słowa. Jedną z propozycji ulepszenia takiego podejścia jest uporządkowanie wyników według określonej miary. To proste rozszerzenie stało się fundamentem wyszukiwarki Google, która zwracane użytkownikowi rezultaty porządkuje według miary jakości oceny źródła, w którym one wystąpiły[1].



Rys. 4. Budowanie grup podobieństw dla artykułów zawierających przykładowe słowo „jądro".

[1] S. Brin, L. Page, *The anatomy of a large-scale hypertextual Web search engine*, „Computer networks and ISDN systems" 1998, no. 30, s. 107–117.

Inną możliwością jest zamiana ciągłych rezultatów wyszukiwania na strukturę grup, które wykazują pewne podobieństwo. W procesie tym wykorzystywane są metody uczenia nienadzorowanego, czyli klasteryzacji[2], umożliwiającej rozpoznanie w zbiorze elementów grup najbardziej podobnych. Zastosowanie tego podejścia zaprezentowano na rysunku 4. Typowe wyszukiwanie oparte na słowach kluczowych zwraca listę ciągłych rezultatów, co przedstawia prawy panel. Z lewej strony zbudowane są grupy opisujące pewne koncepcyjne podobieństwo artykułów zawierających zadane słowo. Grupy te zorganizowane zostały w postaci hierarchii kategorii, które pozwalają na przeglądanie przez ich pryzmat zbioru dokumentów wyznaczonych słowem kluczowym. Na rysunku 4. przedstawiono przykład zorganizowania stron zawierających wieloznaczne słowo „jądro", dla których zbudowane zostały grupy określające różne dziedziny, w których to pojęcie może występować.

4. Podsumowanie i dalsze kierunki rozwoju

W artykule przedstawiono trzy podejścia do graficznej prezentacji Wikipedii umożliwiające nawigację po tym repozytorium wiedzy z użyciem interaktywnych komponentów. Przedstawione zostało również podejście wykorzystujące metody uczenia maszynowego, pozwalające w automatyczny sposób organizować dokumenty w hierarchie tematyczne. Przedstawione w tym artykule aplikacje dostępne są na stronie projektu pod adresem http://swn.eti.pg.gda.pl.

Zaprezentowane metody stanowią w chwili obecnej zalążek architektury systemu, który cały czas jest rozwijany. Metody, które w tej chwili wykorzystujemy, bazują na statycznych metrykach podobieństwa. Planujemy rozwinąć je w kierunku metod umożliwiających definiowanie użytkownikowi, jakiego rodzaju podobieństwa go interesują, co pozwoli w różnorodny sposób organizować i przeszukiwać repozytorium wiedzy[3].

Planujemy również wykorzystać bardziej zaawansowane metody graficznego przedstawiania danych. Do prezentacji większej liczby elementów chcemy użyć hierarchicznych map

[2] R. Xu, D. C. Wunsch, *Clustering*, Wiley-IEEE Press 2009.

[3] J. Szymanski, W. Duch, *Dynamic Semantic Visual Information Management*, w: *Series of Information and Management Sciences, California Polytechnic State University, 9th Int Conf on Information and Management Sciences (IMS 2010)*, s. 130–113.

samoorganizujących[4]. Podejście oparte na mapach Kohonena[5] pozwoli na prezentacje topologicznego sąsiedztwa obiektów występujących w repozytorium. Umożliwi to przedstawienie złożonych struktur koncepcyjnych, co skutkować będzie szybszym dostępem do treści zgromadzonych w repozytorium.

[4] M. Dittenbach, D. Merkl, A. Rauber, *The growing hierarchical self-organizing map*, w: *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks*, 2000, s. 15–19.

[5] T. Kohonen, *The self-organizing map*, „Proceedings of the IEEE" 1990, no. 78 (9), s. 1464–1480.