STUDIA INFORMATICA 2011
Volume 32 Number 2B (97)

Bartosz DUBEL, Paweł KASPROWSKI
Politechnika Śląska, Instytut Informatyki

AGREGACJA DANYCH TEKSTOWYCH NA PRZYKŁADZIE SYSTEMU INFORMACJI PRASOWEJ

Streszczenie. Nadmiar informacji dostępnej w postaci tekstowej w sieci Internet staje się coraz większym problemem, ponieważ automatyczna analiza takich danych jest trudna. Typowymi przykładami dużych baz tekstowych są serwisy prezentujące bieżące informacje prasowe. Z uwagi na dużą liczbę takich serwisów, wiele informacji powtarza się. W artykule omówiono system z grupy tak zwanych agregatorów, który gromadzi w jednym miejscu informacje z wielu serwisów, dokonuje ich analizy i klasyfikacji, a następnie generuje na ich podstawie różnego rodzaju statystyki.

Słowa kluczowe: dane tekstowe, agregacja, analiza tekstu

AGGREGATION OF TEXTUAL DATA ON EXAMPLE OF PRESS INFORMATION SYSTEM

Summary. Huge amount of textual information available in Internet becomes one of the most important problems because analysis of such data is difficult automatically. Typical examples of such big text databases are web services presenting press information. The same or very similar information repeats in different services. That is why so called “aggregators” that aggregate and preprocess information from different services are becoming more and more popular. This paper presents one of such aggregators that collects information from multiple services, parses and analyses it and then tries to classify and collect different statistics.

Keywords: textual data, text aggregation, text parsing

1. Wprowadzenie

Gwałtowny rozwój Internetu sprawił, iż stał się on źródłem prawie nieograniczonej informacji. Powstały tysiące stron udostępniających liczne informacje na każdy temat. Duża liczba

źródeł informacji jest zaletą, ale i wadą Internetu. Zaletą, ponieważ posiadamy dostęp do wielu różnych, niezależnych źródeł wiadomości, które prezentują odmienne punkty widzenia. Wadą, gdyż objęcie tak olbrzymiej ilości danych przez jedną osobę jest bardzo trudne, a czasami wręcz niemożliwe. Zapoznanie się z codzienną porcją nowości wymaga poświęcenia czasu, zwłaszcza, że wiele informacji jest powielanych przez różne źródła. Problemem jest też uzyskanie konkretnej informacji na temat, który jest przedmiotem naszego zainteresowania.

Sytuacja ta sprawiła, że w Internecie zaczęły pojawiać się serwisy oferujące w jednym miejscu zlepek informacji z różnych źródeł – tak zwane *agregatory newsów*. Są one pewnym ułatwieniem dla osób chcących szybko dowiedzieć się, jakie nowe informacje pojawiły się na ich ulubionych serwisach informacyjnych. Agregatory newsów mają jednak kilka wad – m.in. skróty informacji prezentowane w nich są zwykle pobierane z kanałów RSS/Atom [1]. Zawarte tam informacje są zwykle tylko streszczeniem właściwych artykułów i często dają niewielkie pojęcie o właściwej treści całego opracowania. Kanały te nie są też oferowane przez wszystkie strony. Ponieważ informacje udostępnione w kanałach RSS/Atom są niepełne, znacznie utrudnione, a czasem niemożliwe, jest przeprowadzenie pełnej analizy tekstów i porównanie ich podobieństwa czy rozpoznanie ich tematyki. Sprawia to, że agregatory newsów są jedynie stronami prezentującymi zlepek skrótów artykułów z różnych stron. Nie posiadają możliwości personalizacji prezentowanych informacji ani ich przeszukiwania. Co więcej, często są to skróty artykułów o podobnej czy wręcz takiej samej treści, publikowanych na różnych stronach.

Przedmiotem omawianego w artykule projektu było zaprojektowanie, wykonanie i przeprowadzenie analizy działania Systemu Informacji Prasowej. Jest to system komputerowy wspomagający zbieranie, przetwarzanie i prezentację informacji w sieci Internet. Jego głównym zadaniem jest regularne zbieranie informacji dostępnych na stronach internetowych, głównie o profilu informacyjnym i publicystycznym. Te informacje są następnie wstępnie analizowane i przetwarzane, a potem zapisywane. Analiza i przetwarzanie opierają się głównie na powiązaniu wiadomości, artykułów i innych treści zebranych z wielu stron w sekcje tematyczne i utworzeniu wzajemnych relacji między nimi. Ma to na celu ułatwienie późniejszej prezentacji w formie spersonalizowanej dla indywidualnego użytkownika, a także pozyskanie informacji statystycznych na temat zebranych danych.

Użytkownicy korzystający z serwisu mają dostęp do paneli podzielonych tematycznie. Zawierają one informacje na interesujące ich tematy. Artykuły można przeglądać według różnych kryteriów – czasu publikacji, strony, na której się ukazały czy tematu, jakiego dotyczą. Głównym celem systemu jest prezentacja wiadomości z uniknięciem dublowania tych samych danych. Dodatkowo system udostępnia pewne informacje analityczne – statystyki popularności fraz, możliwość przeszukiwania zebranych danych.

2. Architektura

Główne zadania stawiane przed systemem to analizowanie stron WWW i prezentacja zebranych i przetworzonych informacji użytkownikowi końcowemu w serwisie WWW. Każde z tych zadań wymaga zastosowania nieco innej technologii i innego podejścia. Analiza i przetwarzanie stron internetowych muszą być realizowane przez aplikację działającą nieprzerwanie, według zdefiniowanego harmonogramu. Serwis internetowy powinien działać na serwerze WWW i udostępniać swoją zawartość klientom (użytkownikom) komunikującym się z nim przez sieć Internet. Wymusza to rozbicie systemu na dwie, w dużym stopniu niezależne aplikacje. Dlatego architektura każdej części będzie rozpatrywana osobno. Część pierwsza to robot internetowy (ang. *web crawler*), zbierający informacje ze stron WWW, które następnie są przetwarzane i zapisywane do bazy danych. Druga aplikacja to serwis internetowy, którego głównym zadaniem jest prezentacja zebranych przez robota danych i statystyk utworzonych na ich podstawie.

2.1. Web crawler

Web crawler to program zbierający informacje ze stron internetowych. Robot odwiedza strony w regularnych odstępach czasu i przegląda je w poszukiwaniu interesujących go informacji. Jest to aplikacja działająca samodzielnie, potrzebuje jedynie dostępu do sieci Internet. Z tego względu rozbudowany interfejs użytkownika jest zbędny. Może ona działać w dwóch trybach – jako program konsolowy lub jako demon systemowy. Pierwsza metoda przeznaczona jest do testowania, natomiast druga do zastosowań produkcyjnych. Crawler składa się z sześciu modułów: managera, parsera plików konfiguracyjnych, agregatora, parsera HTML/XHTML, modułu logującego i generatora statystyk.

2.2. Serwis WWW

W serwisie internetowym można wyodrębnić ogólnodostępną część główną i panel administracyjny, dostępny dla uprawnionych użytkowników. Panel administracyjny umożliwia między innymi wygodny dostęp do modeli aplikacji z poziomu interfejsu webowego. Jego architektura jest ściśle oparta na strukturze bazy danych. Część główna – czyli właściwy serwis WWW – to system zarządzania treścią, umożliwiający przeglądanie informacji pobranych i przetworzonych przez web crawler oraz statystyk utworzonych na ich podstawie. Serwis WWW jest tworzony z zastosowaniem wzorca MTV, który przypomina wzorzec MVC. Idea MTV opiera się na rozdziale aplikacji na trzy główne komponenty: model (opis danych i logiki programu), widok (określa, jakie dane mają być prezentowane) i szablony (sposób pre-

zentacji danych). Kontrolerem aplikacji jest framework webowy. Architektura taka zapewnia dużą przejrzystość, łatwość rozbudowy i konserwacji aplikacji.

3. Analiza zawartości serwisów informacyjnych

Zbieranie i prezentacja informacji z różnych źródeł wymaga ujednolicenia ich formatu. Jest to konieczne do osiągnięcia szybkiego indeksowania danych, pozwala na łatwiejsze ich przeszukiwanie i klasyfikację do działów tematycznych. Ważne jest pobranie tekstów w taki sposób, aby zachować ich pierwotną hierarchię, np. przynależność artykułów do kategorii. Serwisy internetowe posiadają stosunkowo zbliżoną hierarchię i budowę swoich stron, dlatego realizacja tego zadania nie jest znaczącym problemem, ale wymaga utworzenia spójnego modelu wspólnej dla wszystkich artykułów bazy.

Jako podstawową jednostkę można wyróżnić artykuł – jest to informacja dotycząca jednego tematu. Jako jego stałe elementy można uznać tytuł i treść. Są to elementy najważniejsze i występujące praktycznie we wszystkich serwisach informacyjnych. Kolejną istotną właściwością artykułu jest czas publikacji, aczkolwiek nie można go uznać za stały element, ponieważ jego poprawny odczyt nie zawsze jest możliwy, nawet, jeśli jest określony (wynika to z różnego sposobu zapisu stosowanego w różnych serwisach). Artykuł często posiada też podtytuł i wstęp, zwykle jedno z tych dwóch pól. Służą one do uwydatnienia dodatkowych, kluczowych dla czytelnika informacji o treści publikacji. Są to elementy opcjonalne, ale ze względu na swój charakter istotne, ponieważ zawierają dużo słów kluczowych, dobrze opisujących sam tekst.

Dodatkowymi elementami pomagającymi sklasyfikować artykuł tematycznie są kategorie i słowa kluczowe. Zarówno tych pierwszych, jak i drugich może być do artykułu przyporządkowanych wiele, ale sposób ich odczytu znacząco się różni. Słowa kluczowe powiązane z publikacją znajdują się na jednej stronie, zawsze w jednym miejscu. Natomiast kategorie, do których przyporządkowany jest artykuł, można odczytać, analizując zawartość stron, które zawierają odnośniki do niego (nazywanych dalej stronami linkującymi) oraz strukturę strony artykułu. Ta druga metoda opiera się na założeniu, że na stronie, na której opublikowany jest artykuł, znajduje się menu z odnośnikami do różnych kategorii. Kategoria, do której należy sam artykuł, jest zwykle dodatkowo oznakowana, np. przez inny kolor przycisku, pogrubienie itp. Jeśli tak jest, możemy określić kategorię danego artykułu przez prostą analizę kodu HTML jego strony.

3.1. Analiza danych tekstowych

Wygląd stron internetowych znacząco ewoluował w ciągu ostatnich lat. Mimo to, obecnie nadal główną technologią wykorzystywaną do prezentacji treści w Internecie, z wykorzystaniem przeglądarek internetowych, jest język HTML. Obecnie odchodzi się od wykorzystywania znaczników HTML do definiowania sposobu prezentacji danych. Zaleca się stosowanie do tego stylów CSS, natomiast HTML powinien służyć jedynie do opisu struktury strony WWW. Język HTML stale ewoluuje, a jego kształt określa m.in. konsorcjum W3C, wyznaczając standardy, które w dużym stopniu są wdrażane przez producentów przeglądarek internetowych [2]. Niestety twórcy stron internetowych często nie stosują się do tych standardów, nadal częste jest wykorzystywanie języka HTML jako nośnika informacji o sposobie wyświetlania strony WWW. Powoduje to trudności w poprawnym odczycie zawartości strony w sposób zautomatyzowany. Niemal niemożliwe jest w pełni automatyczna analiza składniowa stron serwisu i rozdział ich na sekcje, a następnie określenie, która część to tytuł, gdzie się kończy, a gdzie zaczyna właściwa treść. Proces ten dodatkowo utrudniają reklamy czy np. skróty innych artykułów wplecione w tekst. Często dokładną analizę stron utrudniają skrypty Java Script. Implementacja obsługi tego języka w robocie internetowym jest bardzo trudna, a dodatkowo powodowałaby znaczny narzut czasowy podczas analizy (tzw. parsowania).

Powyższe powody czynią w pełni zautomatyzowaną analizę niemożliwą. Istnieje jednak możliwość częściowej mechanizacji procesu odczytu treści ze stron WWW, minimalizując udział człowieka do określenia pewnych parametrów dla każdego serwisu. Dzieje się tak dzięki językowi XPath [3], który umożliwia adresowanie części dokumentu, bazując na modelu DOM [2], wykorzystywanym w dokumentach HTML i XHTML. XPath posiada bardzo duże możliwości „wyłuskiwania” treści z dokumentów. Dlatego też został on wykorzystany w omawianym rozwiązaniu do stworzenia plików konfiguracyjnych XML, opisujących strukturę analizowanych stron. Dzięki językowi XPath możemy określić, w jakim miejscu strony znajduje się tytuł, gdzie się kończy, a gdzie zaczyna treść artykułu.

Taki sposób przekazywania konfiguracji silnikowi analitycznemu narzuca jednak pewne ograniczenia. W przypadku zmian w modelu DOM strony (czyli zmianie jej struktury), konieczne staje się poprawienie istniejącej konfiguracji lub napisanie nowej. Jest to problem praktycznie nie do uniknięcia, jednak na szczęście w praktyce zmiany w strukturze przeciętnego serwisu internetowego są wprowadzane stosunkowo rzadko, zwykle co kilka lat. Nie zmienia to faktu, że struktura plików konfiguracyjnych i kodu źródłowego stron, dla których został stworzony, powinna być co jakiś czas analizowana. W rozpoznawaniu konieczności wprowadzenia zmian w plikach konfiguracyjnych dużą pomocą jest rozbudowany system logujący.

Inną wadą stosowania plików z ustawieniami dla każdego obsługiwanego serwisu jest konieczność tworzenia takich konfiguracji dla każdego serwisu z osobna. Zwykle taki proces nie sprawia problemów, jednak dla niektórych stron wymaga dokładnego debuggowania i testów. Dużą pomocą w tworzeniu konfiguracji są specjalne tryby uruchomieniowe crawlera (m.in. debug, parsowanie szablonów, praca tylko jednego serwisu). Określanie ścieżek XPath jest dużo łatwiejsze z wykorzystaniem dodatków do przeglądarek przeznaczonych dla programistów, np. Firebug (dostępny dla przeglądarek Firefox i Chrome), WebDeveloper (Chrome) oraz XPath Checker (Firefox).

Pewnym problemem podczas tworzenia konfiguracji dla serwisów WWW jest złożoność języka XPath. Daje on duże możliwości, jednak jego składnia sprawia, że czasem jest on mało czytelny. Rozwiązaniem tego problemu są dodatkowe parametry konfiguracyjne w plikach XML, które pozwalają w sposób bardziej czytelny uzyskać taki sam efekt, jak przy wykorzystaniu wyrażeń XPath. Nie ograniczają się one jednak tylko do tego. Dają też możliwość ignorowania treści znajdujących się wewnątrz interesujących nas sekcji (m.in. reklam często zagnieżdżanych w artykułach) oraz wykorzystania wyrażeń regularnych do wyłuskiwania treści.

3.2. Tworzenie kategorii

Klasyfikacja artykułów jest konieczna w celu określenia ich przynależności do kategorii. Kategoria jest to zbiór artykułów powiązanych wspólnym tematem, np. Sport czy Polityka. Do poprawnego sklasyfikowania artykułów do kategorii konieczne jest utworzenie jej definicji, zawierającej listę słów opisujących daną kategorię. Dla Sportu byłaby to lista zawierająca nazwy dyscyplin, ale i pojęcia blisko powiązane z wiadomościami sportowymi, jak np. „trener", „puchar", „liga".

Dobieranie odpowiednich słów wymaga jednak dogłębnej analizy informacji, jakie mają zostać sklasyfikowane, a także testów. Odnosi się to w szczególności do kategorii ogólnych, jak Sport, Polityka, Gospodarka czy Kultura. Przykładowo, oczywiste wydaje się przypisanie do Sportu słowa „euro", jako kojarzącego się z Mistrzostwami Europy w Piłce Nożnej. Słowo „euro" to także waluta wielu państw europejskich i sklasyfikowane mogą zostać także artykuły o tematyce gospodarczej czy o polityce. W tym przypadku możemy rozwiązać problem dla kategorii sportowej, dodając listę słów ignorowanych i umieszczając tam zbiór: [polityka, gospodarka, giełda, inflacja, przyroda, morderstwo].

Kolejnym rozwiązaniem może być zamiana słowa „euro" na frazę „euro 2012". Z jednej strony zawężamy wyszukiwanie do konkretnych Mistrzostw Europy, ale z drugiej eliminujemy możliwość traktowania waluty euro na równi z mistrzostwami w piłce nożnej. Takich problemów jest wiele i są one ściśle związane z językiem. Tworząc kategorię „Konflikty

zbrojne", można by się pokusić o opisanie jej słowami „konflikt", „wojna", „atak". Jednakże są to słowa bardzo często używane do opisu wydarzeń sportowych czy politycznych.

Stworzenie odpowiednich definicji dla kategorii nie jest zadaniem łatwym i jednorazowym. Konieczna jest regularna ich analiza i wprowadzanie poprawek uwzględniających najnowsze wydarzenia.

3.3. Przyporządkowywanie artykułów

Każdy artykuł składa się z kilku sekcji. Możemy zwykle wyodrębnić tytuł, podtytuł, wprowadzenie, właściwą treść, słowa kluczowe i kategorię. Słowa kluczowe i kategoria są sekcjami, które można w najprostszy sposób wykorzystać. Ich przeszukiwanie pod kątem zgodności do słów kluczowych kategorii zdefiniowanych w systemie praktycznie zawsze ma sens i nie powoduje błędnej klasyfikacji, ponieważ autorzy tych tekstów w serwisach źródłowych zwykle dbają o odpowiedni dobór słów kluczowych i poprawne dobranie kategorii. Eliminuje to możliwość złej klasyfikacji. Można by zatem postawić pytanie, czy klasyfikując artykuły, można korzystać jedynie z tych dwóch sekcji? Niestety nie zawsze. Pierwszym problemem jest to, że słowa kluczowe, choć często wykorzystywane w wielu serwisach informacyjnych, nie są stosowane wszędzie. Druga kwestia to mocno ograniczona informacja, jaka znajduje się w kategoriach i słowach kluczowych. Jest to praktycznie tylko kilka słów, co prawda bardzo dobrze opisujących sam artykuł, jednak uniemożliwiających głębszą analizę i dokładniejszą klasyfikację. Przykładowo, klasyfikacja do kategorii Polityka na podstawie kategorii pobranej ze strony źródłowej, daje bardzo dobry efekt w większości przypadków, ponieważ w prawie każdym serwisie jest taka kategoria. Gdyby jednak interesowały nas informacje na temat polityki jednego państwa, np. „polityka Niemiec", wtedy poprawna klasyfikacja na podstawie słów kluczowych i kategorii strony źródłowej staje się niemożliwa. W takim przypadku konieczne staje się wykorzystanie informacji zawartych w pozostałych sekcjach artykułu.

Analiza pełnej treści artykułu, czyli jego tytułu, podtytułu, wprowadzenia i treści właściwej, daje olbrzymie możliwości, jednak nakłada też pewne ograniczenia. Przede wszystkim konieczne jest użycie bardziej wyrafinowanych metod przeszukiwania tekstu i pogodzenie się z wolniejszą klasyfikacją z powodu znacznego zwiększenia się ilości tekstu, który należy przeszukać. W przypadku analizy kilku sekcji, istotne jest też nietraktowanie ich równorzędnie. Znalezienie słów kluczowych w tytule czy podtytule jest istotniejsze niż w treści. Jest to spowodowane tym, że tytuł jest zazwyczaj dobierany w taki sposób, aby jak najlepiej ukazywało, o czym jest cały artykuł. W samej treści mogą się znajdować słowa mające wiele znaczeń i na podstawie analizy samej treści artykuł może zostać sklasyfikowany do kategorii całkowicie niepowiązanej z punktu widzenia człowieka. Jest to spowodowane możliwą wielo-

znacznością wielu wyrazów, występowaniem synonimów czy nawet przypadkowym dołączeniem do analizy, jako części artykułu, treści reklam.

W obu metodach klasyfikacyjnych istotne jest określenie stopnia podobieństwa artykułów do definicji kategorii (czyli listy słów ją opisujących). Nie jest to proste ze względu na to, że dla każdego zestawu słów współczynnik podobieństwa dla tego samego artykułu przyjmuje nieco inne wartości. Wymaga to przeprowadzenia testów i dokładnej analizy ich wyników. Dobranie odpowiedniego współczynnika jest tym trudniejsze, im bardziej złożona jest definicja kategorii.

Dodatkowym usprawnieniem, pozwalającym na lepszą klasyfikację, jest możliwość określenia serwisów źródłowych dla kategorii. W przypadku przydzielenia tylko jednego serwisu, kategoria może zawierać artykuły opublikowane tylko w tym serwisie. Daje to możliwość zabezpieczenia się przez przypadkowym sklasyfikowaniem do tej kategorii artykułów z serwisów, które na pewno nie zajmują się daną tematyką.

3.4. Obliczanie podobieństwa tekstów

Porównanie podobieństwa artykułów pozwala na ich lepszą klasyfikację, umożliwia proponowanie użytkownikom informacji podobnych do tych, które właśnie przeglądają, pozwala też odnajdywać publikacje poświęcone temu samemu tematowi, wydarzeniu lub osobie. Porównywanie takie daje bardzo duże możliwości, jednak ma pewne ograniczenia. Algorytmy porównujące dłuższe teksty są mocno obciążające dla bazy danych. Największym problemem przy analizie licznego zbioru artykułów jest złożoność takich porównań, która rośnie wykładniczo. W praktyce oznacza to, że gdy liczba tekstów, które mają zostać porównane, dojdzie do pewnego pułapu (zależnego od wydajności sprzętowej i algorytmu używanego do porównań), porównania stają się nieopłacalne ze względu na długi czas ich wykonania. Aby zwiększyć wydajność tej operacji, w generatorze statystyk obsługa porównań tekstów została całkowicie przerzucona na bazę danych. Dzięki wykorzystaniu procedur składowanych bazy danych nie jest konieczna dodatkowa komunikacja z bazą danych, dzięki czemu wydajność znacząco wzrasta. Do procedur przekazywane są identyfikatory obiektów, które mają zostać porównane, a następnie, po wykonaniu porównań, wyniki zapisywane są do osobnej tabeli w bazie danych. Jest to możliwe dzięki wykorzystaniu biblioteki dodatkowej pg_trgm bazy PostgreSQL, która pozwala określać podobieństwa pomiędzy dwoma tekstami. Korzysta ona z metody porównywania trigramów.

Ponieważ głównym problemem przy porównywaniu tekstów jest rosnąca wykładniczo złożoność, należy starać się maksymalnie ograniczyć liczbę porównań. Można tego dokonać, wprowadzając wstępną selekcję, która przed właściwą operacją porównania odrzuci te artykuły, które z dużym prawdopodobieństwem i tak nie wykazałyby zadowalającego stopnia po-

dobieństwa do pozostałych tekstów. W praktyce do realizacji tego zadania można wykorzystać klasyfikację artykułów do kategorii. Po przeprowadzeniu tej klasyfikacji (która znacznie mniej obciąża bazę danych dzięki wykorzystaniu silnika wyszukiwania pełnotekstowego Sphinx), artykuły zostają podzielone na mniejsze grupy. Jako drugie kryterium ograniczające ilość porównań artykułów, można wykorzystać czas. Zwykle artykuły w dużych serwisach informacyjnych pojawiają się bardzo szybko, dlatego można przyjąć, że czas pomiędzy ukazaniem się informacji o tym samym wydarzeniu nie przekroczy jednego, dwóch dni. Kryterium to nie tylko przyczynia się do wzrostu wydajności systemu, ale i jakości. Dzieje się tak, gdyż zmniejsza się ryzyko uznania za takie same dwóch informacji, które, mimo iż bardzo podobne lingwistycznie, dotyczyły dwóch różnych wydarzeń.

Opisywane porównywanie artykułów tak naprawdę odnosi się do dwóch rodzajów porównań – dla tytułów i dla treści publikacji. Porównywane są tylko te dwie sekcje, ponieważ każdy artykuł w bazie je posiada, a poza tym są one najbardziej miarodajne. Istotnym czynnikiem jest określenie minimalnego stopnia podobieństwa, przy którym dwa artykuły zostaną uznane za podobne. Współczynnik ten należy określić osobno dla porównań tytułów i porównań zawartości tekstów. Najlepsze rezultaty osiągane były, kiedy jego wartość dla tytułów wynosiła 0,3, a dla treści 0,36 (gdzie: 1 oznacza dwa identyczne teksty, a 0 oznacza dwa teksty, zupełnie niepodobne).

3.5. Statystyki

Generacja statystyk to analiza pobranych z serwisów WWW danych pod względem popularności pewnych fraz. Mogą być tworzone na podstawie zdefiniowanych słowników zawierających frazy wzorcowe lub w sposób w pełni automatyczny. Pierwsza metoda znajduje zastosowanie zwłaszcza przy tworzeniu rankingów popularności konkretnych osób (np. polityków, sportowców) czy wydarzeń. Druga powinna dawać bardziej ogólne wyniki. Po przeprowadzeniu testów obu metod okazało się, iż jedynie pierwszy sposób daje zadowalające wyniki. Możliwe jest dzięki niemu uzyskanie miarodajnych rankingów dla wybranych fraz. Problematyczne okazały się jednie różne nazwy stosowane w odniesieniu do tych samych osób, wydarzeń czy przedmiotów. Problem ten udało się rozwiązać, stosując dla wyszukiwanych fraz słownik dostępnych wariacji, przykładowo dla frazy „Janusz Korwin-Mikke” dostępnymi wariacjami mogą być „Korwin-Mikke” i „Korwin Mikke”. Wykorzystanie wariacji znacząco podnosi skuteczność generacji takich statystyk.

Implementacja drugiej – w pełni automatycznej – metody opierała się na wyszukiwaniu najpopularniejszych słów przy użyciu wbudowanych mechanizmów bazy PostgreSQL (funkcja ts_stat). W praktyce okazało się, że nawet po odrzuceniu z wyników tak zwanych *stop words* (czyli nieznaczących części mowy – przyimków, spójników itd.), udział nieznaczących

czasowników był zbyt duży, aby osiągnąć zadowalający efekt. Aby w praktyce wykorzystać tę metodę, należałoby stworzyć dłuższą listę słów pomijanych, co jest jednak trudne do osiągnięcia w sposób arbitralny i uniwersalny dla każdej kategorii.

Rys. 1. Slot rankingowy – ranking zdefiniowanych fraz
Fig. 1. Ranking slot for defined phrases

4. Analiza rezultatów

W celu sprawdzenia skuteczności systemu przeprowadzone zostały testy, w czasie których crawler zbierał informacje publikowane w dziewięciu serwisach internetowych przez okres pięciu dni. Wybrane zostały popularne portale, na których dziennie jest publikowanych od kilku do kilkudziesięciu artykułów na różne tematy:

- www.dziennik.pl
- www.forbes.pl
- www.gazeta.pl
- www.gazetaprawna.pl
- www.money.pl
- www.rp.pl
- www.tvn24.pl
- www.tvp.info
- wiadomosci.polska.pl
- wiadomosci.wp.pl
- wyborcza.pl

Dla celów klasyfikacji utworzone zostały 23 kategorie testowe, na podstawie których zostały przyporządkowane artykuły. Dla każdej z nich minimalny współczynnik zgodności ustawiony został na bardzo niskim poziomie (0,1). Utworzone kategorie można podzielić na ogólne (np. Świat, Polska) i dotyczące konkretnych zagadnień (np. Euro 2012, Zakajew).

- Świat [świat | zagranica] LP: NIE, AZ: NIE

- Polska [polska | kraj] LP: NIE, AZ: NIE
- Polska popularne [polska | zagranica] LP: TAK, AZ: NIE
- Rosja [rosja | putin | miedwiediew | moskwa | petersburg | 'federacja rosyjska'] LP: NIE, AZ: TAK
- Niemcy [niemcy | berlin | merkel | frankfurt | bonn | Hamburg] LP: NIE, AZ: TAK
- Biznes [biznes] LP: NIE, AZ: NIE
- Biznes popularne [biznes] LP: TAK, AZ: NIE
- Gospodarka [(giełda | =akcje | obligacja | 'rada polityki pieniężnej' | 'stopa procentowa' | vat | podatek | in_acja | gospodarka | wig | nasdaq | bank | gpw | inwestycja | inwestor | waluta | emerytura | ofe | reforma | bezrobocie| finanse | inwestycja | pieniądze | dolar | kredyt | ubezpieczenie | fundusz | ofe | rynek | 'wall street' | budżet | kryzys | hossa | bessa | bankrut | przedsiębiorca | firma | ekonomia | ekonomista) –sport] LP: NIE, AZ: TAK
- Gospodarka popularne [gospodarka] LP: TAK, AZ: NIE
- Kultura [kultura] LP: NIE, AZ: NIE
- Nauka [nauka | wiedza] LP: NIE, AZ: NIE
- Pogoda [pogoda | meteo] LP: NIE, AZ: NIE
- Sport 1 [(medal | lekkoatletyka | judo | rekordzista | kibic | sport | sportowiec | trener | trening | łyżwy | piłkarz | 'piłka nożna' | koszykówka | 'skoki narciarskie' | siatkówka | tenis | łyżwiarstwo | liga | pzpn | euro | mistrzostwa | puchar | liga | zawody | mistrzostwa | puchar | sezon | 'rzut młotem' |narciarstwo | kolarstwo | boks | karate | 'piłka ręczna' | małysz | formuła | rajd | hokej | stadion | boisko | kort | bieżnia | futbol | kubica | gortat | otylia | sparing | mecz | piłka | atletyka) & -bank & -giełda & -polityka & -sld &-pis] LP: NIE, AZ: TAK
- Sport 2 [sport] LP: NIE, AZ: NIE
- Polityka 1 [sejm | senat | parlament | rząd | minister | 'unia europejska' | ue | usa | 'stany zjednoczone' | prezydent | irak | afganistan | wojna | rosja | smoleńsk | 'katastrofa smoleńska' | poseł | senator | premier | wicepremier | kanclerz | sekretarz | dyplomacja | komisja | afera | cba | cbś | konflikt | tusk | komorowski | kaczy«ski | palikot | kalisz | olejniczak | maciarewicz | cimoszewicz | buzek | balcerowicz | napieralski | seneszyn | kempa | 'korwinmikke' | lepper | giertych | oleksy | pawlak | kwaśniewski | wałęsa | Sikorski | wasserman | 'marek jurek' | 'ludwik dorn' | gosiewski | kamiński | miller | krrit | abonament | prokuratura | wojsko | grom | partia | media | polityka | polityk | pis | 'prawo i sprawiedliwość' | 'platforma obywatelska' | psl | sld | lpr | socjalista | liberał | konserwatysta | lewica | obama | berlusconi | sarkozy | merkel | putin | miedwiediew | janukowycz | timoszenko | zapatero| 'komisja europejska' | =europejskie | 'gordon brown' | sondaż | pikieta | manifestacja | demonstracja | protest | strajk] LP: NIE, AZ: TAK
- Polityka 2 [polityka] LP: NIE, AZ: NIE

- Polityka popularne [polityka] LP: TAK, AZ: NIE
- Katastrofa Smoleńska [katyń | 'lech kaczyński' | 'pałac prezydencki' | smoleńsk | 'katastrofa smoleńska' | Tupolew] LP: NIE, AZ: TAK
- Euro 2012 ['euro 2012' | 'mistrzostwa w piłce nożnej' | 'mistrzostwa świata w piłce nożnej' | 'polskie stadiony'] LP: NIE, AZ: TAK
- Wybory samorządowe ['kampania samorządowa' | 'wybory samorządowe' | samorząd] LP: NIE, AZ: TAK
- Radio Maryja ['radio maryja' | rydzyk | redemptorysta | 'ojciec dyrektor'] LP: NIE, AZ: TAK
- Zakajew [czeczenia | zakajew] LP: NIE, AZ: TAK

Na liście LP oznacza „łącz podobne" – jeśli ustawione na TAK, system starał się automatycznie pomijać artykuły bardzo podobne do już zaindeksowanego. AZ oznacza natomiast „analiza zawartości" – w kategoriach, gdzie ten parametr był ustawiony na NIE treść artykułu nie była w ogóle analizowana.

Tabela 1

Wyniki klasyfikacji artykułów

Kategoria	Przyporządkowane artykuły			
	Suma	Prawidłowo	Prawidłowo (%)	Złączone
Świat	832	827	99%	-
Polska	869	867	100%	-
Polska popularne	171	171	100%	20
Rosja	264	230	87%	-
Niemcy	144	123	85%	-
Biznes	148	148	100%	-
Biznes popularne	35	35	100%	5
Gospodarka	0	0	-	-
Gospodarka popularne	31	31	100%	0
Kultura	187	186	100%	-
Nauka	134	134	100%	-
Pogoda	6	5	83%	-
Sport1	268	201	75%	-
Sport2	145	145	100%	-
Sport popularne	135	135	100%	20
Polityka1	817	735	90%	-
Polityka2	154	154	100%	-
Polityka popularne	50	50	100%	4
Katastrofa Smoleńska	201	112	56%	-
Euro2012	26	25	96%	-
Wybory samorządowe	100	24	24%	-
RadioMaryja	7	7	100%	-
Zakajew	20	20	100%	-

W ciągu pięciu dni zebrane zostały 5 292 unikalne artykuły należące do 112 kategorii i opisane 1125 słowami kluczowymi. Tabela 1 przedstawia uzyskane rezultaty. Suma przyporządkowanych artykułów to wszystkie artykuły, jakie zostały przydzielone w procesie automatycznej klasyfikacji do kategorii. Spośród nich część została przydzielona błędnie z punktu widzenia czytelnika. Weryfikacja tych przydziałów została przeprowadzona na podstawie subiektywnej oceny.

Najwyższa skuteczność klasyfikacji została osiągnięta dla kategorii, w których odbywała się ona tylko na podstawie analizy słów kluczowych, stron, z jakich pochodziły, i kategorii ze stron źródłowych. Skuteczność klasyfikacji jest wtedy prawie stuprocentowa, jednak zastosowanie tylko tej metody klasyfikacji ma sens jedynie w przypadku bardzo ogólnych i popularnych kategorii, jak Świat, Polska, Sport czy Polityka, ponieważ istnieją one w większości dużych serwisów WWW.

Klasyfikacja z uwzględnieniem analizy zawartości artykułów powoduje przydzielanie do kategorii znacznie większej liczby artykułów, jednak odczuwalnie zwiększa margines błędu podczas klasyfikacji. Jest to dobrze widoczne na przykładzie kategorii Polityka i Sport. Do tych kategorii, oznaczonych numerem 1, artykuły były przydzielane na podstawie pełnej analizy treści i skuteczność jest tu 10 do 25% niższa niż w kategoriach oznaczonych jako 2. Zważywszy jednak na kilkukrotnie większą liczbę przyporządkowanych artykułów w kategoriach 1, trudno uznać którąś z tych metod za jednoznacznie lepszą. Istotnym utrudnieniem przy tworzeniu definicji kategorii, która zostanie wykorzystana przy analizie zawartości tekstów, jest konieczność doboru odpowiednich fraz. Wymaga to poświęcenia pewnej ilości czasu na analizę lingwistyczną publikacji.

Wydaje się, że największą skuteczność system prezentuje dla kategorii specjalistycznych, zwłaszcza dotyczących konkretnych osób i instytucji (np. kategoria dotycząca Ahmeda Zakajewa). Definicje są wtedy proste i krótkie, a w konsekwencji ryzyko błędnej klasyfikacji jest niskie. Kategorie poświęcone państwom (Rosja, Niemcy) nie dały już tak zadowalających rezultatów. Główną przyczyną jest częste używanie fraz z nimi powiązanych w publikacjach w gruncie rzeczy poświęconych innym tematom. Rozwiązaniem tego problemu mogłoby być podniesienie minimalnego współczynnika zgodności, jednak odbiłoby się to negatywnie na liczbie poprawnie przyporządkowanych do kategorii artykułów.

Najgorsze rezultaty otrzymano dla kategorii Gospodarka, Katastrofa Smoleńska i Wybory Samorządowe. Kategoria Gospodarka dobrze obrazuje trudność z dobraniem odpowiedniej definicji. Ta wykorzystana podczas pomiarów wydaje się dobrze opisywać tę kategorię, jednak nie przynosi oczekiwanych rezultatów. Problemem w tym przypadku jest trudność w dobraniu słów kluczowych, opisujących temat gospodarki. Ponieważ obecna definicja całkowicie nie spełnia oczekiwań, należałoby wykorzystać w niej frazy bardziej ogólne. Kategorie

Katastrofa Smoleńska i Wybory Samorządowe, mimo iż posiadają stosunkowo dużo sklasyfikowanych artykułów, to wiele z nich jest błędnie przydzielonych. Wykorzystane w nich słowa kluczowe pojawiają się wielu publikacjach i przez swoją popularność powodują klasyfikację wielu niepożądanych artykułów. Niestety dla podanych kategorii trudno znaleźć alternatywne definicje. Jest to przykład, że klasyfikacja na podstawie samych fraz nie zawsze daje zadowalające rezultaty.

Powyższe badania zostały przeprowadzone z wykorzystaniem identycznego współczynnika zgodności dla wszystkich kategorii. Był on na tyle niski, że praktycznie akceptował wszystkie artykuły zgodne z definicją kategorii. W celu zwiększenia skuteczności klasyfikacji i zmniejszenia liczby błędnie przyporządkowanych artykułów, należałoby podnieść wartość współczynnika. Takie działanie ma jednak negatywny wpływ na liczbę klasyfikowanych artykułów. Generalnie im wyższy współczynnik, tym mniej artykułów jest klasyfikowanych do kategorii, jednak wyższy jest poziom ich zgodności z kategorią.

5. Podsumowanie

Przeprowadzone testy pokazały skuteczność aplikacji w zbieraniu i przetwarzaniu danych uzyskanych z sieci Internet. W zależności od kategorii, dla której klasyfikowane były dane, różna była skuteczność automatycznego przydzielania artykułów. Najlepsze rezultaty zostały osiągnięte podczas zbierania informacji na bardziej wyspecjalizowane tematy, dla których można w sposób jasny określić frazy kluczowe wykorzystywane podczas wyszukiwania. Kategorie bardziej ogólne są trudniejsze do analizy, ze względu na znacznie większy zasób słownictwa z nimi powiązany. Możliwość określenia przedziału czasowego, z jakiego artykuły mają być analizowane dla danej kategorii, dodatkowo podnosi skuteczność analizy dla wydarzeń, o których informacje pojawiają się tylko przez pewien ograniczony okres.

Ponieważ analiza tekstów w systemie opiera się na przeszukiwaniu artykułów pod kątem zgodności z pewnymi frazami zdefiniowanymi przez operatorów, więc na nich spoczywa praca związana z analizą zawartości serwisów internetowych. Jest to najważniejsza część aplikacji, ponieważ od dobrej konfiguracji zależy jakość uzyskiwanych wyników. Konieczne jest też prowadzenie regularnego nadzoru i uzupełnianie fraz wykorzystywanych podczas klasyfikacji w celu utrzymania wysokiej skuteczności działania aplikacji.

Podczas tworzenia aplikacji konieczne było przeprowadzenie dogłębnej analizy wielu stron internetowych i ich kodu źródłowego. Mimo iż sieć Internet podlega nieustannej standaryzacji i coraz popularniejsze są standardy definiowane przez konsorcjum W3C, nadal wiele serwisów ich nie przestrzega. Odległa jest wizja semantycznej sieci, w której dane zawarte w Internecie byłyby zrozumiałe dla programów automatycznie je analizujących. Trudność w od-

czycie informacji ze stron WWW wymusiła użycie plików konfiguracyjnych, zdefiniowanych osobno dla każdego serwisu. Takie rozwiązanie wymusza większą kontrolę człowieka nad procesem pobierania i klasyfikacji artykułów, jednak pozwala też uzyskać lepsze, bardziej wiarygodne wyniki.

Łączone artykuły, czyli dopasowywanie publikacji o bardzo podobnym zasobie lingwistycznym, pozwalają znaleźć teksty opisujące te same wydarzenia. Umożliwia to określenie, jakie wydarzenia są najpopularniejsze, a w konsekwencji najważniejsze. Druga zaleta tej metody to prezentacja użytkownikom tej samej treści tylko raz. Dzięki temu czytelnik nie musi wielokrotnie czytać tej samej informacji prasowej, co ma miejsce w sytuacji, gdy przegląda się wszystkie informacje z kilku serwisów po kolei.

Oprócz standardowej funkcjonalności aplikacja oferuje dodatkowe moduły, m.in. rankingowy, pozwalający na tworzenie i prezentowanie rankingów popularności fraz. Jest to narzędzie bardzo przydatne, zwłaszcza do sprawdzania, jakie osoby, produkty czy instytucje były popularne w ostatnich dniach czy tygodniach. Innym usprawnieniem jest prezentacja podobnych artykułów. Są to sugerowane teksty o podobnej zawartości do aktualnie przeglądanych przez użytkownika. Wyświetlanie podobnych artykułów jest dobrą metodą na dalsze zatrzymanie czytelnika w serwisie, gdyż stale otrzymuje on informacje, którymi jest zainteresowany.

BIBLIOGRAFIA

1. RSS (Really Simple Sindication), http://www.wikipedia.pl/wiki/RSS.
2. Specyfikacja języka XPath, http://www.w3.org/TR/xpath/.
3. World Wide Web Consortium, http://www.w3.org.
4. Salton G.: Developments in Automatic Text Retrieval. Science. Vol. 253, s. 974-979.
5. Sholom W., White B., Apte C.: Lightweight Document Clustering. IBM T.J. Watsan Research Center, 2000.
6. Kłopotek M. A.: Inteligentne wyszukiwarki internetowe. Akademicka Oficyna Wydawnicza EXIT, Warszawa 2001.
7. Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K., Harshman R.: Indexing By Latent Semantic Analysis. Journal of the American Society For Information Science, Vol. 41, 1990.

Recenzenci: Dr inż. Małgorzata Bach
Dr hab. inż. Krzysztof Goczyła, prof. Pol. Gdańskiej

Wpłynęło do Redakcji 31 stycznia 2011 r.

Abstract

Huge amount of textual information available in Internet becomes one of the most important problems because analysis of such data is difficult automatically. Typical examples of such big text databases are web services presenting press information. The same or very similar information repeats in different services. That is why so called "aggregators" that aggregate and preprocess information from different services are becoming more and more popular. This paper presents one of such aggregators that collects information from multiple services, parses and analyses it and then tries to classify and collect different statistics.

Adresy

Bartosz DUBEL: Politechnika Śląska, Instytut Informatyki, ul. Akademicka 16,
44-100 Gliwice, Polska.
Paweł KASPROWSKI: Politechnika Śląska, Instytut Informatyki, ul. Akademicka 16,
44-100 Gliwice, Polska, kasprowski@polsl.pl.