INSTYTUT INŻYNIERII I GOSPODARKI WODNEJ

POLITECHNIKA KRAKOWSKA im. TADEUSZA KOŚCIUSZKI

Szymon Krywus

SEARCH ENGINES PERSONALISER – SYSTEM ANALIZY I PERSONALIZACJI WYSZUKIWAREK INTERNETOWYCH

praca magisterska

studia dzienne

kierunek studiów: **informatyka**

specjalność: **informatyka stosowana w inżynierii środowiska**

promotor: **dr inż. Robert Szczepanek**

nr pracy: **2182**

KRAKÓW 2008

ul. Warszawska 24, 31-155 Kraków tel/fax (+48 12) 628 20 41 e-mail: sekretariat@iigw.pl internet: www.iigw.pl

Składam serdeczne podziękowania

mojej Mamie i bliskim

za trud włożony w moją edukację oraz wsparcie,

Panu dr inż. Robertowi Szczepankowi

za poświęcony mi czas, wskazówki oraz pomoc przy pisaniu pracy.

1. Cel pracy

Celem tej pracy jest zaprojektowanie i stworzenie systemu o nazwie „Search Engines Personaliser”, a następnie analiza jego działania. System jest narzędziem służącym do personalizacji wyników wyszukiwania oraz analizy wyszukiwarek internetowych pod kątem ich dostosowywania się do potrzeb użytkownika. System oraz przeprowadzona analiza mają posłużyć zbadaniu popularnych i ogólnie dostępnych wyszukiwarek internetowych i odpowiedzi na pytanie czy wyszukiwarki internetowe potrafią dostosowywać rezultaty zapytań do potrzeb i oczekiwań użytkowników.

Obszarem nauki który posłuży w przeprowadzonej próbie będzie inżynieria środowiska. Wyszukiwarki zostaną zbadane przy pomocy wybranych wcześniej fraz, które będą zarówno ogólne jak i bardziej szczegółowe, jednak powiązane ze sobą tematycznie. System SEP sprawdzi kolejno frazy w badanych wyszukiwarkach a rezultaty wyszukiwarek zostaną ocenione przez użytkownika. Dzięki zbudowanemu modelowi analizy, system SEP będzie narzędziem którego efekty pracy staną się subiektywną oceną wyszukiwarek internetowych oraz być może próbą odpowiedzi na pytanie, która z badanych wyszukiwarek najlepiej potrafi dostosowywać rezultaty wyszukiwań do konkretnych oczekiwać oraz potrzeb użytkownika.

2. Zakres pracy

Pierwszym etapem było zapoznanie sie z problematyka wyszukiwarek internetowych oraz ich sposobem pracy od strony technicznej. Przybliżając sposoby i technikę działania wyszukiwarek internetowych warto przedstawić w tym miejscu kilka podstawowych pojęć związanych z tą tematyka co będzie pomocne w zrozumieniu tematu.

Po zapoznaniu sie ze specyfikacją i właściwościami wyszukiwarek należało wybrać kilka z nich które były brane pod uwagę przy tworzeniu systemy SEP oraz zostały wykorzystane do analizy jego działania.

Do badania zostały wybrany trzy wyszukiwarki: Google.pl, Netsprint.pl, Szukacz.pl.

Dzięki wybraniu tych wyszukiwarek możliwa była szersza ocena uzyskanych wyników, a co za tym idzie większe prawdopodobieństwo uzyskania lepszych rezultatów podczas analizy. Wybrane wyszukiwarki nie są trzema najpopularniejszymi wyszukiwarkami w Polsce, gdyż według badań te najpopularniejsze należą zazwyczaj do dużych portali internetowych (np. Onet.pl). Jednak takie wyszukiwarki oparte są w większości o system wyszukujący utworzony przez Google. Aby więc skorzystać ze znanych wyszukiwarek ale jednocześnie takich, które różnią się znacząco podstawowymi aspektami takimi jak: silnik wyszukujący, baza wiedzy, algorytmy wyszukujące, wybrane zostały trzy wyszukiwarki: Google.pl, Netsprint.pl, Szukacz.pl. Dzięki wybraniu systemów opartych na różnych technologiach, można zbadać różnice miedzy nimi i spróbować stwierdzić który z nich najbardziej spełnia oczekiwania wyszukiwarki dostosowanej do potrzeb użytkownika.

W pracy opisane zostały także techniki programistyczne użyte do implementacji systemu SEP oraz model bazy danych wykorzystany w systemie.

Następnie przedstawiony został szczegółowy opis poszczególnych elementów systemu SEP oraz algorytmu „multiwyszukiwarki” – zasadniczego moduły systemu.

Kolejne rozdziały poświęcone są działaniu systemu wraz z opisem każdego etapu analizy wyszukiwarek internetowych dostępnych w systemie SEP. Taki przedstawienie kolejnych kroków analizy ułatwiło zrozumienie działania systemu.

Następnie system został przetestowany oraz stworzona została analiza wyszukiwarek internetowych w oparciu o tematyką inżynierii środowiska. W celu stworzenia analizy porównawczej poszczególnym wyszukiwarek, potrzebne było wybranie siedmiu fraz. Frazy były zasadniczą częścią badania wyszukiwarek. Ich kolejność oraz budowa miały

wpływ na wyniki analizy. Frazy zostały kolejno wprowadzone do systemu SEP który pobierał wyniki wyszukiwania z wybranych wyszukiwarek. Kolejno po każdej badanej frazie wyszukiwarki były oceniane. Dzięki ciągłości badania możliwa była końcowa analiza wyszukiwarek na podstawie wszystkich zbadanych fraz.

Na podstawie uzyskanych wyników, powstał końcowy wniosek zawierający ocenę działania systemu SEP oraz wybranych wyszukiwarek.

3. Wyszukiwarki internetowe

3.1 Ogólny opis działania

„Wyszukiwanie informacji to proces wyszukiwania w pewnym zbiorze tych wszystkich dokumentów, które poświęcone są wskazanemu w kwerendzie tematowi (przedmiotowi) lub zawierają niezbędne dla użytkownika fakty i informacje.” [Kłopotek, 2001].

Wyszukiwarka internetowa jest to system (program/strona WWW) służąca do wyszukiwania w zasobach internetu informacji na podstawie określonych kryteriów. Patrząc od strony użytkownika, wyszukiwarki to rozbudowane narzędzia umożliwiające wyszukanie różnego rodzaju dokumentów, zarówno testowych jaki i dźwiękowych czy obrazu. Natomiast od strony programistycznej, jest to zespół oprogramowania gromadzący, segregujący, grupujący informacje o zasobach znajdujących się w sieci.

3.2 Zastosowanie

Wyszukiwarki internetowe stały się w ostatnich latach jednym z najbardziej popularnych narzędzi do odnajdywania informacji w sieci. Ogromny postęp, jaki nastąpił w rozwoju technologii internetowej, jego powszechny dostęp wpłynęło na zwiększenie ilości informacji umieszczonych w sieci. Dzięki temu internet a przede wszystkim wyszukiwarki internetowe, oprócz swojego podstawowego zadania czyli filtrowania i sortowania zasobów internetu, stały się również narzędziem marketingowym. Wyższa pozycji w wyszukiwarkach, a tym samym większa szansa na zauważenie przez potencjalnego użytkownika, skłoniły specjalistów do tworzenia co raz to nowszych i bardziej skutecznych metod pozycjonowania stron. SEO (*Search engine optimization*) to działania zmierzające do wypromowania danego serwisu, strony internetowej na jak najwyższą pozycję w wyszukiwarkach.

Dzisiaj wyszukiwarki internetowe tworzone są przez ogromne korporacje zarabiające miliony dolarów na zyskach jakie przynosi im ich popularność. Dzięki popularności internetu i jego ogromnej ekspansji na rynek marketingowy, wyszukiwarki internetowe

stały się jedną z najbardziej popularnych form reklamy. Możliwość zaistnienia w sieci, pokazania się i zaprezentowania swojej oferty, staje się motorem napędowych dla firm starających się dotrzeć do klienta.

Dziedzina wyszukiwarek wciąż bardzo się rozwija. Przez ostatnich kilka lat wiele z wyszukiwarek przestało istnieć a ich miejsce zajęły kolejne. Jednym z problemów wyszukiwarek jest brak radykalnej, rewolucyjnej zmiany w technologii algorytmów wyszukiwania, języków zapytań, interfejsu użytkownika czy też brak zastosowania algorytmów genetycznych.

3.3 Rodzaje

Globalna sieć internet, to źródło ogromnej ilości informacji. Bogactwo potrzebnej i ciekawej wiedzy niesie ze sobą jednak kilka zasadniczych problemów a główny z nich to trudność do ich dotarcia. Szacuje się, że obecna ilość stron WWW przekracza liczbę 185 milionów *[źródło http://news.netcraft.com/].* Dlatego też w gąszczu informacji płynących z internetu ciężko znaleźć tę, która satysfakcjonowałaby potencjalnego użytkownika.

Wraz z upływem czasu i rozwojem technologii internetowej, pojawiły się coraz to skuteczniejsze metody wyszukiwania informacji w sieci. Podczas przeglądania internetu można natknąć się na dużą ilość programów wspomagających i ułatwiających wyszukiwanie. Można je podzielić ze względu na sposób działania na następujące kategorie *[Kłopotek, 2001]*.:

1. „Oświecone zgadywanie”
2. Katalogi stron
3. Przewodniki i specjalizowane katalogi przedmiotowe, bazy wiedzy itp.
4. Kolekcje linków, generowane głównie automatycznie (np.: FFA – Free For All Links)
5. Portale, wortale, „strony startowe”

6. Wyszukiwarki (szperacze) indeksujące
7. Metawyszukiwarki
8. Osobiste narzędzia multiwyszukiwawcze (Multi Search desktop software)
9. Inne serwisy wyszukiwawcze (centra wyszukiwacze, wyszukiwania przez ludzi)
10. Rozwiązania mieszane (najpierw poszukiwanie w katalogach, potem zapytania do wyszukiwarki)

Warto w tym miejscu zaznaczyć, iż stworzony system SEP należy do kategorii numer 8.

Systemy wyszukujące można podzielić na części:

1. Pozyskujące dokumenty
2. Indeksujące
3. Wyszukiwawcze
4. Bazy danych
5. Tymczasowe składowiska dokumentów

Dane zapisywane w bazach danych pozyskiwane są automatycznie przez roboty lub przez człowieka.
Systemy wyszukujące w zależności od swoich możliwości technologicznych i zastosowania, mogą pozyskiwać ogromne ilości informacji, te mniejsze natomiast skupione tylko na danej dziedzinie proporcjonalnie mniej.

3.3.1 Oświecone zgadywanie

Technika ta polega na zgadywaniu adresów URL przy pomocy poszukiwanego słowa/frazy. I tak chcąc znaleźć informacje na temat opadu atmosferycznego możemy

skonstruować adres http://opad-atmosferyczny.pl itp. Zaletą takiej metody jest szybkość i intuicyjność jednak zasadniczą wadą jest bardzo niska skuteczność. Przy prostych frazach i słowach ma ona swoje zastosowanie, natomiast przy próbie znalezienia odpowiedzi na skomplikowane, rozbudowane pytanie, praktycznie jest bezskuteczna.

3.3.2 Katalogi stron

Metoda polegająca na przeszukiwaniu „katalogów sieciowych", które udostępniają swoją bazę danych wraz z opisami stron WWW.

Katalogi mają strukturę drzewa. Kategorie dzieląc się na pod kategorię a te jeszcze dalej tworzą zhierarchizowaną strukturę tematyczną. Zagłębiając się w drzewo kategorii jesteśmy w stanie znaleźć interesujące nas strony WWW z danej dziedziny.

Przypisywaniem stron WWW do danej gałęzi katalogu, zajmują się redaktorzy. Ich zadaniem jest przeglądanie poszczególnych stron i w zależności o treści i tematyki przypisanie ich bądź też nie do danej kategorii. Przy dodawaniu strony do bazy danych redaktorzy opisują krótko stronę w celu łatwiejszego znalezienia jej przez użytkowników. Redaktorzy to najczęściej pracownicy firm zajmujących się katalogowaniem stron bądź też ochotnicy. Istnieją także specjalne formularze dzięki których właściciele stron WWW mogą sami katalogować swoje strony. Powstaje w ten sposób uporządkowane drzewko kategorii w postaci hipertekstowej, gdzie na końcu tej struktury znajdują się odsyłacze do stron WWW.

Zaletą katalogów stron jest łatwa i zrozumiała struktura kategorii oraz to, że jest ona tworzona przez człowieka a co za tym idzie prawdopodobieństwo znalezienia odpowiedzi na pytanie jest dużo większe. Jednak ta zaleta może być też wadą, gdyż redaktorzy przypisują strony do kategorii wedle swojego subiektywnego widzenia, co może powodować problemy u użytkowników których spojrzenie na poszukiwana tematykę jest inny.

Największą wadą katalogów stron jest ich aktualność. W czasach kiedy ilość stron gwałtownie rośnie, ich ilość przerasta często możliwości sukcesywnego uaktualniania zawartości katalogów. Aktualizacja lub usunięcie danej strony z zasobów sieciowych nie pociąga za sobą odpowiednich zmian w strukturze drzewa katalogu.
Dlatego też czynnik ten wpływa na to, że kataloguje się tylko niewielko ilość stron. Największe zasoby katalogowe należą do takich gigantów jak Yahoo czy Google.

3.3.3 Przewodniki i specjalizowane katalogi

Katalogi tego typu to odpowiedź na trudny do pominięcia problem aktualizacji katalogów sieciowych. Przewodniki tematyczne, specjalizowane katalogi, bazy wiedzy to wąskie tematycznie katalogi zasobów sieciowych przeznaczone dla wąskiej grupy ludzi interesujących się daną dziedziną. Nie są to komercyjne, rozbudowane katalogi a to z tej przyczyny, iż im bardziej specjalistyczna dziedzina tym mniej potencjalnych użytkowników odwiedza takie katalogi. Dlatego też katalogi tego typu zazwyczaj prowadzone są przez hobbystów, lecz brak wsparcia ze strony sponsorów spycha je na dalszy plan w dziedzinie wyszukiwania informacji w internecie.

3.3.4 Portale, wortale

Portale czy wortale to komercyjne strony, tworzone z myślą o dużej ilości użytkowników. Chęć przyciągnięcia większej ilości użytkowników determinuje ich twórców to publikacji jak najbardziej popularnych informacji. Horoskop, wiadomości, pogoda, darmowe konta pocztowe, czat room-y, fora to tylko mały procent całości jaki składa się na portale. W dzisiejszych czasach tego typu strony coraz bardziej poszerzają swoją ofertę dla klientów. Tworzone są portale 24 godzinne, które to przy pomocy

najnowszych technik multimedialnych zaczynają wypierać popularne metody serwisów informacyjnych. Przykład: Onet.pl, Gazeta.pl, TVN24.pl.

3.3.5 Wyszukiwarki

Wyszukiwarki internetowe od dłuższego czasu są najlepszą metodą wyszukiwania informacji w sieci. Wpływa na to kilka czynników, takich jak duży zakres poszukiwań, szybkość oraz duża dokładność. Wyszukiwarki internetowe odnajdują adresy stron WWW pasujące do podanego przez użytkownika słowa lub frazy. Wyszukiwarka przeszukuje swoją bazę danych a następnie wyświetla rezultaty w postaci uporządkowanej listy adresów stron WWW. Zazwyczaj lista wzbogacana jest o krótki opis, informacje o dokumencie a czasami także zrzut ekranu wyszukanej strony w odpowiednim pomniejszeniu.

Forma zapytania jaką może wpisać użytkownik to zazwyczaj pojedyncze słowo lub kilka slow kluczowych. Wyszukiwarki obsługują jednak także bardziej rozbudowane zapytania. Mogą to być złożone wyrażenia logiczne a nawet pytania w języku naturalnym. Podejmowane są też próby głosowych komend dla wyszukiwarek, jednak w tej kwestii pozostaje kilka problemów takich jak np. ograniczona ilość słów rozumianych bez nauki przez komputer.

Wyszukiwarki składają się z rozbudowanych modułów oprogramowania, które głownie zajmuje się zbieraniem, segregowaniem i udostępnianiem dokumentów z sieci. Oprogramowanie zajmujące się zbieraniem dokumentów to tak zwane roboty, pająki (ang. robot, spider, crawler). Ich zadaniem jest odwiedzanie stron WWW, analiza znajdującej się na nich treści oraz krążenie do następnych stron, których adresy znajdzie. Kolejnym zadaniem robota jest uaktualnianie informacji o stronach, czyli systematyczne odwiedzanie ich i sprawdzanie czy nie zmieniła się ich treść.

Kolejnym elementem wyszukiwarki internetowej jest baza danych. Baza uzupełniana jest automatycznie. Indeksy bazy tworzone są przez oprogramowanie zwane indekserem. Indekser próbuje samodzielnie określić do jakiej kategorii tematycznych

zaliczyć daną stronę i jego wartość, co ma następnie swoje odzwierciedlenie w kolejności jakiej układana jest lista odpowiedzi. Poprzez wybór słów kluczowych w przeglądanym dokumencie, który odbywa się na podstawie analizy położenia poszczególnych słów i ilości ich powtórzeń w stosunku do innych, tworzone jest streszczenie przeglądanego przez indekser dokumentu. Wszystkie te zabiegi prowadzą do stworzenia struktury która umożliwia szybki dostęp do informacji podczas korzystania użytkownik z wyszukiwarki.

Korzyści jakie płyną z wykorzystania wyszukiwarek internetowych daje im znaczną przewagę nad katalogami. Podstawowy aspekt to szybkość. Użytkownik po wpisaniu wyrażenia otrzymuje informacje w czasie poniżej 1-ej sekundy. Jest to nieporównywalne z tym co oferują katalogi, czyli czasochłonne przeszukiwanie kolejnych gałęzi drzewa. Jedna przewaga jaka istniała po stronie katalogów czyli opis recenzji strony pisany przez człowieka, powoli też zanika gdyż firmy utrzymujące katalogi nie są w stanie aktualizować zawartości internetu proporcjonalnie do wzrostu liczby stron WWW. Przez to pogarsza się jakość wyników oferowanych przez katalogi. Natomiast rozwój technologii i algorytmów wyszukiwania, korzystanie z coraz to większych i aktualniejszych baz danych wpływa na rosnącą popularność wyszukiwarek internetowych.

3.3.6 Multiwyszukiwarki (metawyszukiwarki)

Ze względu na duże zróżnicowanie pod względem znajdywanych adresów stron przez różne wyszukiwarki dla danej frazy, zaczęto tworzyć multiwyszukiwarki. Multiwyszukiwarki są to narzędzia które adresują pytanie do kilku wyszukiwarek naraz, a wyniki które otrzymują łączą tworząc mniej lub bardziej jednorodną listę. Multiwyszukiwarki tak naprawdę niewiele różnią się od zwykłych wyszukiwarek internetowych. Podstawowa różnica to taka, że multiwyszukiwarki nie posiadają własnej bazy danych, lecz korzystają z innych.

Multiwyszukiwarki można podzielić na *[Kłopotek, 2001]*:

1. „Lista”
2. Poszukujące pojedynczo
3. Poszukujące równolegle

Serwis typu „lista” to po prostu strona WWW na której zgromadzone zostały odnośniki do wyszukiwarek. Znajdują się tam często odnośniki do bardzo wielu wyszukiwarek, niekoniecznie użytecznych.

Serwis poszukujący pojedynczo udostępnia zazwyczaj kilka najbardziej popularnych wyszukiwarek z których można wybrać te, w których chcemy wyszukać daną informacje. Słowo kluczowe wpisujemy w pole i zaznaczając odpowiednie pola wybieramy interesujące nas wyszukiwarki. Wadą tego typu multiwyszukiwarek jest to, że dane z poszczególnych wyszukiwarek ściągane są jedna po drugiej, co znacznie wydłuża czas oczekiwania na rezultaty.

Serwisy poszukujące równoległe to multiwyszukiwarki w pełnym tego słowa znaczeniu. Są one szybkie, gdyż pobierane wyniki z poszczególnych wyszukiwarek odbywają się w tym samym czasie. Dodatkowo większość z tego typy multiwyszukiwarek korzysta z kilkudziesięciu a nawet kilkuset wyszukiwarek, co znacznie zwiększa ilość otrzymywanych rezultatów.

3.4 „Inteligentne” wyszukiwarki

Wraz z rozwojem technologii informatycznym i systemów internetowych a także zasobów internetowych, rośnie także potrzeba coraz lepszego udoskonalania wyszukiwarek internetowych. Oczekiwania użytkowników są coraz większe. Rośnie potrzeba maksymalnego uproszczenia interfejsu obsługi wyszukiwarek z równoczesnym wzrostem ich skuteczności.

Przeciętny użytkownik nie zapoznaje się z instrukcją tworzenia poprawnych i bardziej efektywnych zapytań podawanych w wyszukiwarkach. Często są one niesprecyzowane,

ogólne co powoduje znaczny wzrost otrzymywanych rezultatów, niekoniecznie relewantnych. Jednocześnie użytkownik oczekuje, że już na pierwszej stronie listy rezultatów znajdzie stronę o treści go zadawalającej. Prowadzi to do potrzeby pomocy użytkownikowi, podpowiedzi jasnej i zrozumiałej ze strony systemy wyszukującego. Jasnej i zrozumiałej aby mógł w jak najszybszym czasie znaleźć interesujące go dokumenty, adekwatne to jego oczekiwań.

Wszystko to skłania twórców wyszukiwarek do opracowywania coraz to nowszych, skuteczniejszych programów które będą „myśleć" za użytkownika. Jednym słowem „inteligentnych". Dlatego też dąży się do tego aby większość obowiązków przy wyszukiwaniu przejęły „inteligentne" systemy wyszukuje. Mają one informować o istotności i wadze dokumentu dla potrzeb użytkownika, to znaczy umieszczać krótkie opisy, słowa kluczowe itp.. Uczyć się profilu użytkownika i pod tym kątem kreować wyniki wyszukiwań. Podpowiadać słowa kluczowe, formy fraz przy pomocy których użytkownik otrzyma prawdopodobnie bardziej korzystne wyniki.

Spełnienie tych wymagać pociąga za sobą konieczność prowadzania w systemach wyszukujących modeli sztucznej inteligencji, opierających się na technologii przetwarzania i rozumienia tekstu w języku naturalnym.

3.5 Budowa wyszukiwarki

3.5.1 Pająk

Pająk jest to program lub grupa programów przeszukująca i ściągająca z internetu jak największą ilość dokumentów, plików które zawierają odpowiednią i przydatną treść, na serwer wyszukiwarki.

Pająki mogą być budowane o powstałe już i dostępne elementy. Są one zaprogramowane w języku Perl bądź w Java, jak np.:

- Checkbot (Unix, Windows; perl5)
- Fluid Dynamics SEPrch Engine robot (Unix, Windows; perl5)
- SpiderMan (Java).

Pająk przeszukując sieć WWW w poszukiwaniu dokumentów, tworzy kolejkę adresów URL. Z każdej napotkanej strony, pająk „wyciąga" linki i dokłada to kolejki linków oczekujących. Wraz z coraz głębszym przeszukiwaniem stron, kolejka linków staje się coraz dłuższa. Ogromna ilość stron WWW jakie napotyka na swojej drodze pająk wpływają na czas jego pracy. Obecne pająki są w stanie przetworzyć nawet ponad 25 stron na sekundę, co i tak powoduje kilkudniowy czas oczekiwania na odświeżenie strony.

Jednak zaawansowane pająki zostały rozbudowane o wiele dodatkowych funkcji, które zapewniają szybsze przeszukiwanie stron WWW. Umiejętność zarządzania pamięcią, procesorem oraz natężeniem ruchu w sieci wpływają na znaczną poprawę szybkości ich działania. Aby zmniejszać ilość wejść na strony znajdujące się na tym samym serwerze, pająki tworzą kolejki i odwiedzają np. tylko pierwsze adresy z każdej z nich co sprawia, że żaden z serwerów nie jest przeciążany, a co za tym idzie poprawia się praca samego pająka.

Jeżeli można podzielić pająki pod względem ich metodologii działania, to tworzyłyby one trzy grupy: pająki błądzące, pająki posługujące się opisanymi już drogami, pająki mieszane. Pierwszy z typów posługuje się początkowo niewielka listą adresów stron WWW, następnie poprzez znalezione na danych stronach adresy wędruje dalej. Drugi typ to pająki które przeszukując inne bazy danych wyciągając z nich potrzebne informacje. Niezbędne są do tego programy przeszukujące inne serwisy. Trzeci typ to pająki korzystające z połączonych dwóch wcześniejszych metod, czyli najpierw odbywa się odnajdywanie potrzebnych informacji, to jest adresów URL z innych serwisów, a następnie przeszukiwanie ich pod kątem adresów do innych stron WWW.

Praca pająka ma przede wszystkim dostarczyć jak najwięcej informacji o danej stronie WWW. Jednak obecne coraz doskonalsze pająki muszą także przewidywać czy dana strona może być interesująca, czy treści na niej zawarte będą przydatne przed jej ściągnięciem do swojej bazy danych. Dlatego też inteligentny pająk opierając się o dostępne mu informacje o danym dokumencie jest wstanie określić jego przydatność, a tym samym pobrać go lub odrzucić. Pająk może znaleźć informacje o danym adresie

WWW w swojej lub innej bazie danych (jeżeli posiada dostęp do takowej). Może również odwołać się do dokumentów które posiadały link do danej strony i poszukać opisów tych linków, które zazwyczaj zawierają krótka charakterystykę strony na którą kierują. Jeszcze innym sposobem pająka jest dokonanie analizy adresu URL danej strony pod kątem nazwy pliku, serwera.

Po wstępnej fazie akceptacji danej strony WWW, pająk musi przeprowadzić ocenę przydatności treści strony. Pomocą w takiej sytuacji staje się dla pająka baza wiedzy która posiada. Z jednej strony spełnia ona rolę bazę lingwistycznych haseł, pod kątem których przeszukiwana jest treść strony. Jeżeli w dokumencie znajdą się poszukiwane hasła kojarzone są one ze słowami kluczowymi których szukamy. Taką rolę spełniają np. tezaurus, czyli ontologiczna baza danych. Z drugiej strony należy posługiwać się pewnymi określonym regułami, kierującymi sposobem doboru dokumentów. Reguły te powinny opisywać postępowanie z danym dokumentem. Jeżeli na przykład dokument zawiera wyrazy należące do pewnego zbioru wyrazów S, to ten dokument zawiera również linki do stron które mogą nas zainteresować. Lub też jeżeli dokument zawiera wyrazy należące do pewnego zbioru S, to ten dokument zawiera również informacje, treści nas interesujące. Natomiast patrząc z innej strony, jeżeli dokument jest główną stroną firmy zajmującej się hydrologią, to strona taka zawiera linki do stron o podobnej tematyce. „Myśląc" w ten sposób, pająk tworzy zagłębione drzewo linków powiązanych ze sobą stron. Im głębiej porusza się on po tym drzewie, tym prawdopodobieństwo natrafienia na relewantne dokumenty jest mniejsza.

Po przeprowadzonej klasyfikacji dokumentów przez pająka, pojawia się zasadniczy problem jego działania czyli sposób sortowania sprawdzonych dokumentów w taki sposób, aby te najbardziej relewantne znalazły się na samej górze listy rezultatów. Dzięki temu, że w bazie danych pająków znajdują się słowa kluczowe wraz z przyporządkowanymi im wagami, możliwa jest klasyfikacja adresów URL dokumentów na podstawie sumy wag słów kluczowych znajdujących się w opisach tych dokumentów. Bazy danych pająków zazwyczaj są tworzone automatycznie. Stosuje się do tego kilka metod, takich jak reguły JEŻELI-TO, sieci neuronowe, algorytmy genetyczne.

3.5.2 Indekser

Indekser to jeden z najważniejszych elementów systemu wyszukiwarki. To oprogramowanie odpowiedzialne za indeksowanie dokumentów znalezionych przez pająka. Indeksy w bazie danych powstałe w procesie indeksowania muszą być zoptymalizowane, tak aby dostęp do nich był szybki i pozwalał na łatwe znalezienie dokumentów odpowiadających podanym słowom kluczowym.
Sposób działania indeksera można podzielić na kilka etapów *[Kłopotek, 2001]*:

1. Identyfikacja słów/fraz/termów występujących w dokumentach
2. Usuwanie słów popularnych
3. Ekstrakcja tematów słów przy użyciu szukającego tematu
4. Zastąpienie tematów przez numeryczne identyfikatory termów indeksujących (w celu wydajnego przetwarzania)
5. Zliczanie wystąpień tematów (obliczanie tzw. tf – term frequency)
6. Opcjonalnie użycie tezaurusa dla termów o niskiej częstości (zastąpienie terminem ogólniejszym)
7. Opcjonalnie tworzenie fraz dla termów o wysokiej częstotliwości
8. Obliczanie wag dla wszystkich prostych termów, fraz i klas tezaurusa (w oparciu o stosowany później model wyszukiwania)
9. Przypisanie każdemu dokumentowi przynależnych prostych termów, fraz i klas tezaurusowych z odpowiednimi wagami.

Jednym z ważniejszych elementów indeksera jest analizator. Jest to narzędzie odpowiedzialne za podział dokumentu na „termy” czyli frazy, słowa, wyrazy. Analizator przeglądając treść dokumentu, rozbija ją na małe elementy (frazy). I tak na przykład zdanie „To są czerwone swetry” analizator może podzielić na „to” „są” „czerwone” „swetry”, zmieniając wielkości liter na małe. Inny analizator może przekształcić frazy na liczbę pojedynczą i pomijając pospolite wyrazy „to” i „są”, tworząc: „kolorowy” „sweter”. Słowa pospolite najczęściej niebrane pod uwagę. Pomija się je biorąc pod uwagę tzw. stop-lista. Taka stop-lista jest różna w zależności od języka a także od rodzaju dokumentów w jakich ma zostać zastosowana. W języku angielskim mogą ją

tworzyć takie wyrazy jak: „the”, „a”, „an”, „at”, „in”, „will”, „there”, „with” itd. Natomiast dla dokumentów o znaczeniu prawnym, stop-listę mogą tworzyć np. „paragraph” itp. Analizatory posiadają również funkcje dopasowywania znalezionych fraz do tematu. Oznacza to, że na przykład angielskie słowa „writen”, „writing”, „writer” zastąpią jednym słowem „write”. Innym przykładem jest rozbijanie rozbudowanych wyrazów na mniejsze, tak jak na przykład słowo: „dezoksyrybonukleinowy” analizator może rozbić na termy: „dezoksy”, „rybonukleinowy”. Analizator stosuje także uogólnienia, synonimy które zwiększają szanse na uzyskanie większej poprawności wyników. I tak analizator może zmienić piesek, bernardyn, jamnik na pies. Oczywiście potrzebna jest do tego odpowiednia ontologiczna baza wiedzy dla określonego języka.

Tak wygenerowane termy są jeszcze analizowane pod kątem, ilości wystąpień w dokumencie, miejsca ich wystąpienia (tytuł dokumentu, nagłówki), ważność termu na tle całego dokumentu. Dzięki takiej rozbudowanej analizie możliwe jest opracowanie streszczenia dla badanej strony, przypisania jej słów kluczowych i kategoryzacji.

4. Narzędzia i metody

4.1 PHP/PHP5

4.1.1 PHP

PHP to skryptowy język programowania, przeznaczony do tworzenia rozbudowanych aplikacji na stronach WWW. Programy w języku PHP do działania potrzebują serwera PHP który wykonuje polecenia zawarte w programie a następnie wysyła odpowiedź do użytkownika. Język PHP mimo swojej prostoty zawiera dużo przydatnych przy tworzeniu dynamicznych stron WWW właściwości. Ma on wbudowane funkcji obsługi połączeń z bazą danych co sprawia, że programy napisane w tym języku umożliwiają operacje na dużej ilości danych. Przy pomocy PHP możemy dynamicznie tworzyć obrazy w różnych formatach, takich jak .jpeg czy .gif.

Pliki PHP mają rozszerzenie .php. Umieszczane są w kodzie pomiędzy znacznikami *<?php (treść skryptu) ?>.*

Najprostszym przykładem działania PHP jest wyświetlenie informacji w oknie przeglądarki. Służy do tego polecenie *echo*.

Każda instrukcja w PHP musi być zakończona średnikiem w przeciwnym wypadku skrypt zwróci błąd.

Komentarze w języku PHP mogą być tworzone dwojako. Pierwszy rodzaj komentarza, tworzony za pomocą dwóch znaków *//* powoduje, że znaki za znakiem komentarza aż do znaku końca lini są ignorowane. Ten sam efekt uzyskamy stosując znaki */** i **/*. Znak */** pokazuje początek komentarza natomiast znaki **/* pokazuje jego koniec. Druga metoda jest bardziej przydatna przy rozbudowanych wielo liniowych programach.

W PHP wprowadzono znaczne uproszczenia w operowaniu zmiennymi. Nie trzeba ich deklarować, co znacznie upraszcza programowanie w porównaniu z innymi językami np. C/C++. Typ zmiennej jest określany automatycznie na podstawie wartości jaka jest jej przypisana. Każdą nazwę zmiennych poprzedza się znakiem *$ (dolar)*. PHP obsługuje najpopularniejszych typów zmiennych tj.:

$jeden = 7; (integer)

$dwa = 3.14; (double)

$trzy = „magister”; (string)

$cztery = array(); (array) .

Aby w tekście dodać znak specjalny taki jak $(dolar), /(slash), należy go poprzedzić znakiem \ (backslash). Znak nowej lini dodajemy poprzez znaki \n. Konutacja odbywa się poprzez kropkę np.:

$c = ”a” . ”b”;

Kolejnym ważnym aspektem języka PHP jest obsługa tablic. Tablice czyli uporządkowany zbiór elementów, których miejsce przyporządkowane jest według indeksów. Wartości do tablic przypisywane są podobnie jak do zwykłych zmiennych. Kolejność w tablicach jest liczona od 0. Tablice mogą być jedno lub wielo wymiarowe. Na każdej zmiennej, która jest liczbą można w PHP wykonywać rozmaite działania takie jak przypisanie, dodawanie, odejmowanie, mnożenie, dzielenie.

Oprócz działań arytmetycznych, możliwe jest również tworzenie działań logicznych. Operatory logiczne stosowane są w celu określenia relacji między zmiennymi.

PHP udostępnia szereg funkcji wbudowanych, takich jak np.: *echo()*, służąca do drukowania wyników pracy programu na ekranie. Jednak język PHP służy głównie do tworzenia własnych, nowych funkcji, dostosowanych do potrzeb projektu. Każda nową funkcje tworzymy od słowa **function**. Następnie piszemy nazwę nowej funkcji zakończoną nawiasem okrągłym (). W środku nawiasu, po przecinku, listę parametrów. Często zdarza się, że dana funkcja ma wykonać pewne obliczenia a następnie zwrócić wynik. Aby tego dokonać należy posłużyć się poleceniem **return**.

Trudno sobie wyobrazić w dzisiejszych czasach języka programowania, który nie umożliwiałaby współpracy z bazami danych. PHP dzięki wbudowanym funkcją daje takie możliwości. Jedną z baz danych z którą PHP współpracuje jest MySQL. Zasada działania jest bardzo prosta i opiera się na utworzeniu połączenia z bazą danych za pomocą polecenia *mysql_connect(”adres serwera”, ”nazwa użytkownika”, ”hasło”);* .

Funkcja *mysql_connect()* zwraca identyfikator połączenia w przypadku powodzenia, lub **FALSE** w przypadku wystąpienia błędu.

Następnie uaktywniana jest wybrana baza danych za pomocą polecenia *mysql_select_db(”sz_db”, $db);*.

Jeżeli proces ten przebiegnie pomyślnie możemy podjąć próbę wysłania zapytania do bazy. Istniej kilka podstawowych zapytań:

1. Wyciąganie danych z tabeli:

   *mysql_query(”SELECT *FROM nazwa_tabeli WHERE warunek1”);*

2. Dodawanie danych do tabeli:

   mysql_query(”INSERT INTO nazwa_tabeli VALUES(wartość1, wartość2,…) ”);

3. Usuwanie danych z tabeli:

   mysql_query(”DELETE FROM nazwa_tabeli WHERE warunek1”);

Oczywiście to tylko mały zarys możliwości obsługi bazy danych z poziomu języka PHP. Jak widać na powyższym przykładzie, każde polecenie kierowane do bazy danych obsługiwane jest przez funkcje *mysql_query()*. W pierwszym z przypadków zwracana jest tablica rezultatów takiego zapytania. Jeżeli napisalibyśmy ją w taki sposób:

*$odpowiedz = mysql_query(”SELECT *FROM nazwa_tabeli”);*

I jeżeli podana tabela nie byłaby pusta, to do tablicy *$odpowiedz* zwrócone zostałyby wszystkie wartości z tabeli.

4.1.2 PHP5 – programowanie obiektowe

Wraz z rozwojem technologii informatycznych, rozwinęły się również języki programowania. Rozwój technik programowania obiektowego i odkrycie tym samym ich ogromnych zalet, spowodowało jego ekspansję niemal we wszystkich językach programowania. Rewolucja nie ominęła również języka PHP. Wraz z wprowadzeniem wersji 5.0, język stał się w pełni obiektowy.

Obiektowość w programowaniu niesie ze sobą mnóstwo zalet. Jedną z głównych jest *„łatwość przekładania poszczególnych wymogów z obszaru zastosowania na poszczególne moduły kodu”* [PHP5.Programowanie zaawansowane]. Innymi słowy programowanie obiektowe modeluje za pomocą „obiektów”, które często są odpowiednikami z otaczającą nas rzeczywistością, a co za tym idzie jest to podejście bardziej realne i łatwe dla uzmysłowienia. Kolejną ważna zaletą programowania obiektowego jest możliwość wielokrotnego powielania kodu. Dzięki temu zmniejsza się ilość lini kodu, programy stają się bardziej czytelne i zwiększa się ich edytowalność. Często zdarza się, że pisząc aplikację w wielu miejscach używamy tych samych typów danych. Na przykład, w programie który zarządza wypożyczalnia samochodów, będziemy wielokrotnie używać jednej klasy **Samochód**. W wypożyczalni samochodów występuje wiele różnych aut, różniących się od siebie np.: osobowe, ciężarowe. Pisząc jedną klasę która modeluje równymi metodami i właściwościami, otrzymujemy narzędzie które można umieszczać w wieli miejscach w kodzie programu. W proceduralnym podejściu do programowania rozwiązanie takiego problemu zajęłaby wiele lini kodu. Łatwo sobie wyobrazić jakie skutki miałoby wystąpienie błędu w taki programie obsługującym wypożyczalnię samochodów. Jeżeli program operowałby na dużej ilości danych i byłby do tego rozbudowany, znalezienie bledu byłby bardzo czasochłonne. W programowaniu obiektowym, modularność klas zapewnia łatwe odtarcie do błędu i usunięcie go a także możliwość szybkiej rozbudowy danej klasy. Wszystkie właściwości i metody klasy **Samochód** znajdują się w jednym miejscu. Dzięki temu czytelność takiej aplikacji jest o wiele lepsza.

Cały szereg udogodnień jakie niesie ze sobą programowanie obiektowe ujęte jest w kilku jego właściwościach *[PHP5.Techniki zaawansowane]*:

1. Klasy – „wzorce” dla obiektów o kod definiujący właściwości i metody.
2. Obiekty – stworzone egzemplarze klasy, które przechowują wszelkie wewnętrzne dane i informacje o stanie potrzebne dla funkcjonowania aplikacji.
3. Dziedziczenie – możliwość definiowania klas jednego rodzaju jako szczególnego przypadku (podtypu) klasy innego rodzaju (na podobnej zasadzie jak kwadrat określany jest jako szczególny przypadek prostokąta).

4. Polimorfizm – umożliwia zdefiniowanie klasy jako członka więcej niż jednej kategorii klas (tak jak samochód, który jest „czymś, co ma silnik” oraz „czymś, co ma koła”).
5. Interfejsy – stanowią „mowę” na podstawie której obiekt może implementować metodę, nie definiując rzeczywistego sposobu implementacji.
6. Hermetyzacja – możliwość zastrzeżenia dostępu do wewnętrznych danych obiektu.

4.2 MySQL

4.2.1 Wprowadzenie do MySQL

Bazy danych są jednym z najbardziej popularnych systemów do przetwarzania i gromadzenia informacji. Jego początki sięgają lat 60-tych. Jednak trudno sobie wyobrazić bez nich dzisiejsze nowoczesne systemy informatyczne. Dzięki znacznemu zwiększeniu mocy obliczeniowej komputerów, bazy danych jako narzędzie mogące magazynować ogromną ilość danych, stały się ich nieodzownym elementem.
MySQL jest systemem zarządzania bazami danych. Jest on rozwijany przez firmę MySQL AB. Jest oprogramowaniem typu „Open Source”, co oznacza, że każdy może je zainstalować i korzystać nie ponosząc z tego tytułu żadnych kosztów. MySQL ma wiele zalet, które wpływają na jego popularność:

- Napisany w C i C++
- Przetestowany przez szeroką gamę kompilatorów
- Działa na wielu różnych platformach
- Do zapewnienia przydatności wykorzystuje narzędzia GNU Automake, Autoconf i Libtool
- Dostępne są API dla języków C, C++, Eiffel, Java, Perl, PHP, Python, Ruby i Tel

- Pełna wielowątkowość z wykorzystaniem wątków jądra. Może korzystać z wielu procesorów
- Bardzo szybki, oparty na wątkach system alokacji pamięci.

4.2.2 Pojęcia i terminologia baz danych

Podstawowymi pojęciami za zakresu projektowania i modelowania bazami danych są obiekty i relacje. Obiekty są to elementy świata rzeczywistego, których cechy i właściwości umieszczane są w tabelach bazy danych. Dla przykładu tworząc obiekty „student" i „wydział", możemy opisach ich cechy, czyli zapisać w bazie danych studenta który uczy się na wybranym wydziale. Zachodzące powiązanie miedzy tymi obiektami nazywane jest relacją. Czyli student o numerze 1 studiuje na wydziale Inżynierii Środowiska opisuje relacje obiektu „student" do obiektu „wydział".
Istnieje kilka typów relacji:

1. Jeden do jednego
   Przykład: Jednemu studentowi przypisany jest tylko jedno miejsce w akademiku.
2. Jeden do wielu (wielu do jednego)
   Przykład: wielu studentów studiuje na jednym kierunku studiów.
3. Wielu do wielu
   Przykład: Wielu studentów może studiować na kilku kierunkach studiów.

4.2.3 Typy kolumn i danych w MySQL

W MySQL istnieją trzy podstawowe typy kolumn:

1. Liczbowe

2. Łańcuchowe
3. Tekstowe

Oraz typy daty i czasu.

Typ liczbowy jest stosowany do przechowywania liczb. W bazie danych możemy stosować różne rodzaje liczb. Zarówno całkowite (int) jak do zmiennoprzecinkowe (float, double). Typy liczbowe można wyświetlać z zadaną szerokością, natomiast liczby zmiennoprzecinkowe ograniczyć do podanej liczby miejsc po przecinku. Oto przykład:

cena_ksiazki decimal(5,2)

Ograniczenie w powyższym przykładzie wynosi 5 cyfr z 2 cyframi po przecinku.
Omawiając dokładniej typy liczbowe można je podzielić na kilka rodzajów:

1. NUMERIC lub DECIMAL

   Oba typy są identyczne. Wykorzystywane są one do przechowywania dokładnych wartości zmiennoprzecinkowych, takich np. jak wartości pieniężne.
2. INTEGER

   Typ całkowity przechowywany w 4 bajtach. Istnieją także odmiany tego typu, np.: TINYINT (1 bajt), SMALLINT (2 bajty), MEDIUMINT (3 bajty), BIGINT(8 bajtów).
3. FLOAT

   Typ zmiennoprzecinkowy o pojedynczej precyzji. Obejmuje on swoją reprezentacja zakres liczb od $1{,}18 \times 10^{-34}$ do $3{,}40 \times 10^{38}$, podobnie sytuacja wygląda dla liczb ujemnych.
4. DOUBLE

   Typ zmiennoprzecinkowy o podwójnej precyzji. Obejmuje on swoją reprezentacja zakres liczb od $2{,}23 \times 10^{-308}$ do $1{,}80 \times 10^{308}$, podobnie sytuacja wygląda dla liczb ujemnych.

Typy łańcuchowe lub **tekstowe** stosowane są w MySQL w kilku różnych odmianach. Zaliczamy do niego typy:

1. CHAR

   Typ ten przechowuje łańcuchy znakowe o stałej długości. Maksymalna długość CHAR to 255 znaków. W bazie danych określając jakąś kolumnę takim typem zazwyczaj podajemy też jego maksymalną długość, np.:

   student CHAR(100)

   Jeżeli nie określimy długości, system użyje domyślnie CHAR(1). Dane zapisywane w kolumnie typu CHAR zawsze posiadają taką samą, wcześniej zdefiniowaną długość. Niezależnie od długości łańcucha znakowego, długość przechowywanych danych typu CHAR jest uzupełniana spacjami do potrzebnego rozmiaru. Natomiast podczas pobierania danych z kolumny CHAR, spacje są automatycznie usuwane. Przewagą typu CHAR nad innymi typami jest szybkość pobierania danych z kolumn tego typu. Łatwiej jest pobierać dane o takiej samej długości. Jednak odbywa się to kosztem miejsca jakie muszą zająć dane typu CHAR.

2. VARCHAR

   Typ służący do przechowywania łańcuchów znakowych o zmiennej długości. Aby określić szerokość, wykonujemy polecenie:

   student VARCHAR(80)

   podając w nawiasie odpowiednią wartość.

   Zakres wartości to od 0 do 255.

3. TEXT, BLOB

   Typ ten wykorzystywany jest do przechowywania dłuższych fragmentów tekstu. Na tyle długich aby nie można było użyć typu CHAR lub VARCHAR. TEXT i BLOB są do siebie bardzo podobne z tym, że typ BLOB został zaprojektowany do magazynowania danych binarnych a nie tekstu. Oba typy mają swoje odmiany np.: TINYTEXT/TINYBLOB (zakres do 255 znaków/bajtów), TEXT/BLOB (zakres do 65 535 znaków/bajtów), MEDIUMTEXT/MEDIUMBLOB (zakres do 16 777 215 znaków/bajtów), LONGTEXT/LONGBLOB (zakres do 4 294 967 295 znaków/bajtów).

4. ENUM

   Typ, który pozwala na wymienienie zestawu możliwych wartości. Każdy z wierszy może posiadać jedną z wartości znajdującej się w zestawie. Przykład:

student enum(‘m’, ‘k’)

Na podstawie powyższej deklaracji, dla pola student możliwe są wartości ‘m’, ‘k’, NULL lub error.

5. SET

   Typ identyczny co ENUM, z tym wyjątkiem, że wiersze mogą zawierać zestaw wartości z zestawu wyliczenia.

Typy daty i czasu również są obsługiwane przez MySQL. Type:

1. DATE

   Określa datę. MySQL zapisuję datę w standardzie ISO, czyli:

   rok-miesiąc-dzień (RRRR-MM-DD).

2. TIME

   Określa czas wyświetlany, jako:

   GG:MM:SS.

3. DATETIME

   Połączenie typu DATE i TIME w formacie:

   RRRR-MM-DD GG:MM:SS.

4. TIMESTAMP

   Typ przechowujący dane dotycząca daty utworzenia lub ostatniej modyfikacji danego wiersza, jeżeli czas nie została podany lub był ustawiony na NULL.

5. YEAR

   Typ przeznaczony do określania roku.

4.2.4 Modelowanie danych w MySQL

Baza danych to narzędzie do przechowywania i modelowania danymi. MySQL umożliwia rozmaite operacje na danych. Stworzona w odpowiedni, dostosowany do naszych potrzeb, baza danych jest kluczem do efektywnego i szybki pobierania z niej dużej ilości informacji. Często w bardzo rozbudowanych modelu, istnieje potrzeba

pobrania informacji z wielu tabel na raz. Informacji ze sobą ściśle powiązanych. Tworzenie pewnej uporządkowanej listy rezultatów niesie za sobą konieczność modelowania ich kolejnością w zależności od potrzeb. Przy dużej ilości danych, modelowanie nimi na poziomie PHP znacznie spowalnia ten proces, gdyż dane otrzymane z bazy są nieposortowane i dopiero przy pomocy poleceń języka skryptowego odbywa się jego selekcja. Przenosząc znaczną część obowiązku sortowania, grupowania, wybierania potrzebnych informacji na bazę danych, zwiększamy szybkość procesu modelowania danymi na poziomie PHP.

MySQL udostępnia szereg podstawowych zapytań do bazy danych języka SQL. Począwszy od takich poleceń jak: *INSERT, DELETE, UPDATE, REPLACE*.

Aby móc manipulować danymi trzeba je najpierw wstawić do bazy danych. Służy do tego polecenie *INSERT*. Aby zastosować to polecenie należy najpierw wskazać do której tabeli chcemy wstawić dane, *INTO „nazwa tabeli”*, następnie opisujemy dane które chcemy wstawić do bazy danych, *VALUES („wartości dla pól w tabeli”)*. Wartości dla pól należy oddzielać przecinkami. Jeżeli wstawiamy jakąś wartość dla pola z ustawioną opcja *auto_increment,* możemy sami ustalić wartość lub pozwolić aby MySQL zrobił to za nas. Dane typu łańcucha znakowego oraz daty należy ujmować w apostrofy, natomiast dane typu liczbowego nie.

Kiedy chcemy pobrać określone dane z bazy danych stosujemy polecenie *SELECT*. Jest to chyba najważniejsze polecenie MySQL, służące do pobierania wiersza z jednej lub wielu tabel. Zapytanie *SELECT*, tak jak pisałem wcześniej, jest podstawowym i najbardziej rozbudowanym ”narzędziem” do modelowania danym w bazie danych. Dzięki wielu możliwościom ograniczeń wyszukiwań z bazy danych, *SELECT* daje ogromne pole dla znalezienia tylko wybranych, ściśle określonych informacji. Wzorzec ogólny polecenia *SELECT* wygląda następująco:

SELECT kolumny
FROM tabele
[WHERE warunki]
[GROUP BY grupa
HAVING warunki_grupy]]
[ORDER BY sort_kolumn]

[LIMIT ograniczenia];

Oczywiście nie jest to pełnia możliwości zapytania *SELECT*. Istnieje jeszcze co najmniej kilka możliwych operacji.

Opisując kolejno podany wyżej schemat, do wybory określonego wiersza w tabeli spełniającego dane kryteria służy *WHERE*. Podczas tworzenia zapytania do bazy z ograniczeniem *WHERE*, stosujemy operatory takie jak: równości(=), „różny od”(!=, <>), operatory porównania(>, <, <=, >=), operatory *IS NULL* lub *IS NOT NULL* oraz standardowe operatory logiczne(*AND*, *OR*, *NOT*).

Aby zapytanie nie zwróciło nam duplikatów należy użyć opcji *DISTINCT*.

Następną opcją pozwalającą na traktowanie pobranego wiersza jako grupy jest *GROUP BY*. Pozwala ona na grupowanie wyników względem podanych kryteriów. Grupowanie można ułożyć w kolejności rosnącej *ASC* lub malejącej *DESC*. Stosując opcję *GROUP BY* połączoną z *HAVING*, otrzymujemy podobny efekt jak połączenie polecenia *SELECT* z *WHERE*, z tą różnicą, że klauzuli *WHERE* używa się w każdym zapytaniu, w którym sprawdza się warunek dla pojedynczego wiersza, natomiast *HAVING* używa się dla całej grupy.

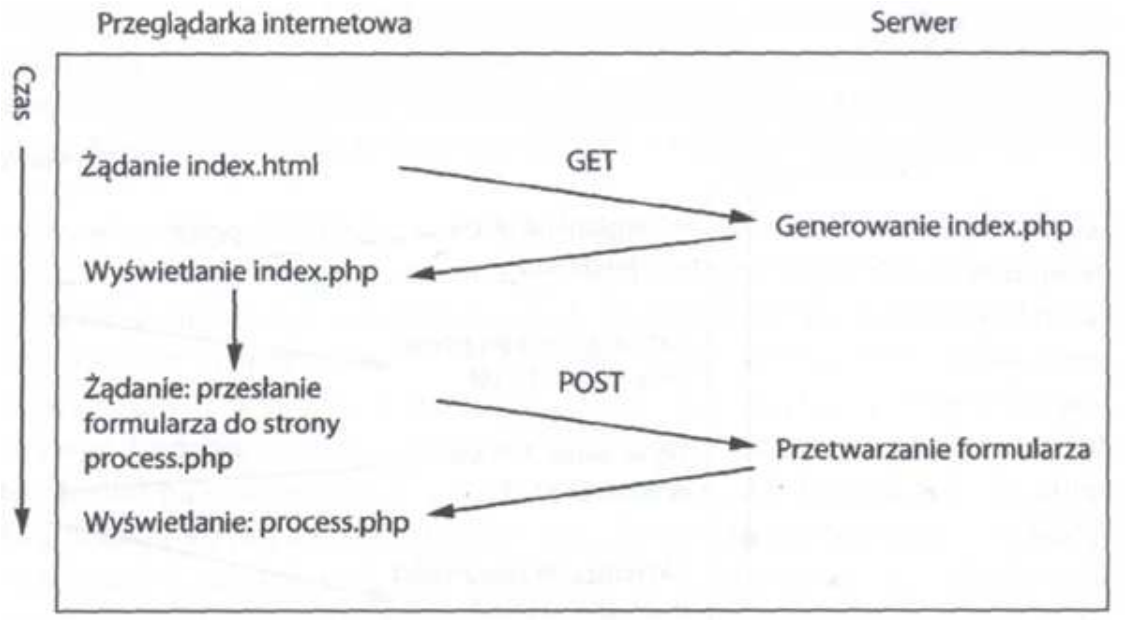
Do sortowania wierszy wyników na podstawie jednej lub więcej kolumn służy opcja *ORDER BY*. Podobnie jak grupowanie, sortowanie może być rosnącej(*ASC*) lub malejące(*DESC*).

Ostatnia opcja to *LIMIT*. Jest ona używana do ograniczenia liczby i zasięgu wierszy zwracanych przez zapytanie. Opcja stosowana najczęściej wraz z klauzulą ORDER BY aby kolejności zwracanych rekordów miała jakiś sens. LIMIT to opcja bardzo przydatna przy stronicowaniu wyników.

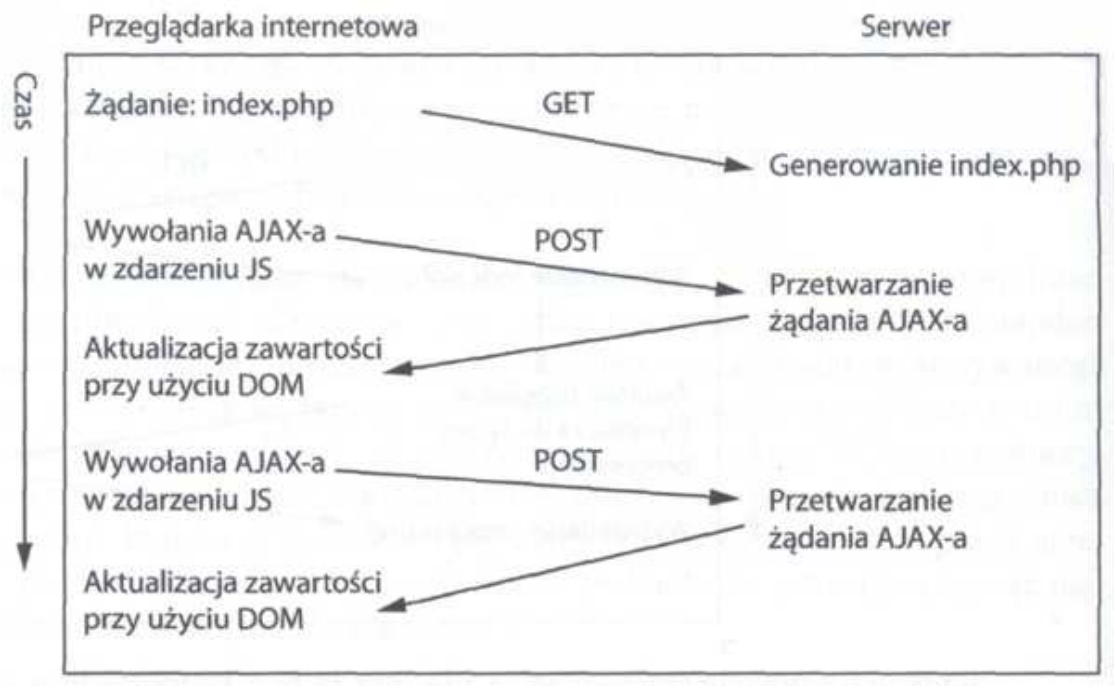
Kiedy w bazie danych znajdują się niepotrzebne dane które należy skasować, posługujemy się wtedy poleceniem *DELETE*. Pozwala ono usunąć wiersze z tabeli. Możemy usuwać wszystkie dane z tabeli, podając jedynie nazwę tabeli, lub także wybrać odpowiednie dane spełniające wskazane przez nas ograniczenia. Stosując polecenie *WHERE*, a następnie wyliczamy ograniczenia które MySQL ma zastosować podczas usuwania danych z tabel.

4.3 AJAX

Termin AJAX (ang. *Asynchronous Java-Script+XML*) pojawił się pierwszy raz w artykule Jesse'go Jamesa Garreta w 2005 roku. Dzięki technologii AJAX możliwy jest nowy sposób komunikacji językiem JavaScript używanym w przeglądarkach internetowych a serwerem. To znacząca zmiana w porównaniu ze zwykłą komunikacja aplikacji WWW a serwerem. Żądania które generowane są w języku AJAX dla serwera nie różnią się niczym od zwykłego zapytania wysyłanego ze strony WWW. Natomiast dla przeglądarki jest to nowa rzecz, ponieważ nie wymagają one odświeżania strony. Różnice pomiędzy zwykłą komunikacją żądanie – serwer oraz żądanie AJAX – serwer widać na rysunkach 1.1 i 1.2.



Rys 1.1. Przepływ żądań w aplikacji internetowej *[„Ajax i JavaScript" Joshua Eichorn]*.



Rys 1.2. Przepływ żądań w aplikacji bazującej na AJAX-ie *[„Ajax i JavaScript" Joshua Eichorn]*.

Przełomem który sprawił że, możliwe było zdefiniowanie języka programowania AJAX było stworzenie przez firmę Microsoft obiektu ***XMLHttpRequest.*** Obiekt ten początkowo miał umożliwić wczytywanie dokumentów XML do kodu JavaScript w przeglądarkach Internet Explorer. XMLHttpRequest to w rzeczywistości klient http używany w języku JavaScript. Dzięki temu możliwe jest tworzenie i wysyłanie na serwer żądań **GET** i **POST**.

Obecnie dostępnych jest dużo bibliotek JavaScript umożliwiające rozbudowane operacje przy użyciu technologii Ajax. Przykładami mogą być tutaj biblioteki:

- protorype
- script.aculo.us
- dojo
- jQuery (wykorzystana w pracy).

4.4 HTML/CSS

4.4.1 HTML

HTML czyli HyperText Markup Language (ang.) został stworzony w oparciu o SGML czyli Standard Generalized Markup Language (ang.). W HTML-u zdefiniowany jest pewien zestaw stylów, używanych do tworzenia stron WWW takie jak: nagłówki, akapity, listy czy tabele. Został również wprowadzony zestaw elementów umożliwiających formatowanie tekstu (np. pogrubienie). Każdy taki element języka HTML posiada swoją nazwę i występuje w formie znacznika. Znacznik HTML nie określa w żaden sposób jak nagłówek czy tabela mają być sformatowane, lecz wskazuje jedynie, że dany element jest nagłówkiem bądź tabelą. Za odpowiednie formatowanie elementów języka HTML odpowiedzialna jest przeglądarka WWW. Pobiera ona całą treść i w odpowiedni sposób rozszyfrowuje elementy na niej zawarte. Przeglądarka posiada wbudowane ustawienia które w dany sposób odzwierciedlają znaczniki HTML na stronie WWW. Standardowe opcje formatowania języka HTML w różnych przeglądarkach różnią się od siebie, co często sprawia wiele problemów przy odpowiednim dopasowaniu strony do każdej z nich.

HTML jest językiem operującym na znacznikach. Oznacza to, że układ i forma strony WWW zależy od użycia odpowiednich, dostosowanych do potrzeb znaczników. Tworząc stronę HTML można operować jedynie na tych znacznikach które są obsługiwane przez przeglądarki. Nie można tworzyć nowych własnych, gdyż przeglądarka nie będzie umiała w odpowiedni sposób wyświetlić ich na ekranie.

Pliki języka HTML to zwykłe pliki tekstowe (ASCII), co oznacza, że nie zawierają one żadnej informacji właściwej dla danej platformy systemowej, co sprawia, że mogą one być odczytywanie praktycznie przez każdy edytor tekstu.

Plik HTML składa się z następujących elementów:

- właściwa treść strony.
- znaczniki HTML, określające elementy strony, jej strukturę, sposoby formatowania i hiperłączenia z innymi stronami bądź z informacjami innego rodzaju.

Znaczniki mają następującą formę:

<znacznikHTML> treść </znacznikHTML>

Praktycznie każdy znacznik składa się z dwóch zasadniczych elementów to jest otwarcia i zamknięcia. Dzięki temu określany jest zakres działania znacznika.

4.4.2 CSS

CSS czyli Cascading Style Sheets (ang.) po polsku tłumaczone jako Kaskadowe Arkusze Stylów, to język służący do zawansowanego opisu formy wyświetlania stron WWW. CSS to zbiór reguł, które ustalają w jaki sposób ma być wyświetlana przez przeglądarkę zawartość wybranych elementów języka HTML. CSS daje szeroki wachlarz możliwości opisu elementów języka HTML takich jak czcionki, tabele, marginesy. Pozwala także na określanie położenia danego elementu względem innych. Zamysłem tworzenia kaskadowych arkuszy stylów była chęć odseparowania od siebie struktury dokumentu i formy jego prezentacji. Taki podziała poprawia czytelność dokumentów a także znacznie ułatwia wprowadzania zmian. Ze względu na to, że arkusze CSS są tworzone zewnętrznie, można jest stosować do wielu plików jednocześnie, bez potrzeby przepisywania stylów dla każdej ze stron osobno.

Arkusz stylów składa się z definicji/opisu stylu dla wybranych elementów strony WWW. Definicja składa się z selektora określającego element lub grupę elementów dla którego tworzone są style. Wartości opisujące daną własność są oddzielone od siebie dwukropkami (:) i zakończone średnikiem (;). Każda definicja zaczyna i kończy się nawiasem klamrowym.

Arkusze stylów są dodawane do strony WWW za pomocą elementu *<link>* lub umieszczane bezpośrednie na stronie za pomocą znaczników *<style></style>*.

Nazwa Kaskadowe Arkusze Stylów wynika z właściwości jaką posiadają. Mianowicie, definicje stylów w arkuszach zewnętrznych, wewnętrznych i na poziomie elementów języka HTML wykluczają się wzajemnie, dlatego też priorytet stylów ustalony jest hierarchicznie. Pierwszeństwo zawsze mają style zdefiniowane bezpośrednio przy elemencie. Później im „dalej" od elementu tym priorytet jest niższy. Kolejność więc wygląda następująco (rosnąco) *[wikipedia.org].*:

1. Domyślny arkusz przeglądarki WWW
2. Domyślny arkusz użytkownika przeglądarki
3. Zewnętrzne arkusze stylów
4. Definicje stylów w nagłówku dokumentu
5. Definicje stylów w atrybucie *style* elementu

5. Funkcjonalny opis systemu SEP

5.1 Techniki implementacji

5.1.1 Użyte technologie

System SEP został stworzony przy użyciu dostępnych darmowych narzędzi do tworzenia programów webowych. SEP oparty jest na technologii HTML/PHP/AJAX, ze względu na dużą dostępności narzędzi to tworzenia tego typu oprogramowania. Oprogramowanie umieszczone na przystosowanym do tego celów serwerze może być używany z każdego komputera na świecie który ma dostęp do internetu i zainstalowaną dowolną przeglądarkę stron WWW.

5.1.2 Konfiguracja systemowa

Do stworzenia systemu SEP użyte zostały następujące technologie informatyczne:

1. Serwer Apache wersja 2.0.50
2. PHP wersja 5.2.6
3. MySQL wersja 5.0.51 a
4. phpMyAdmin wersja 2.11.6

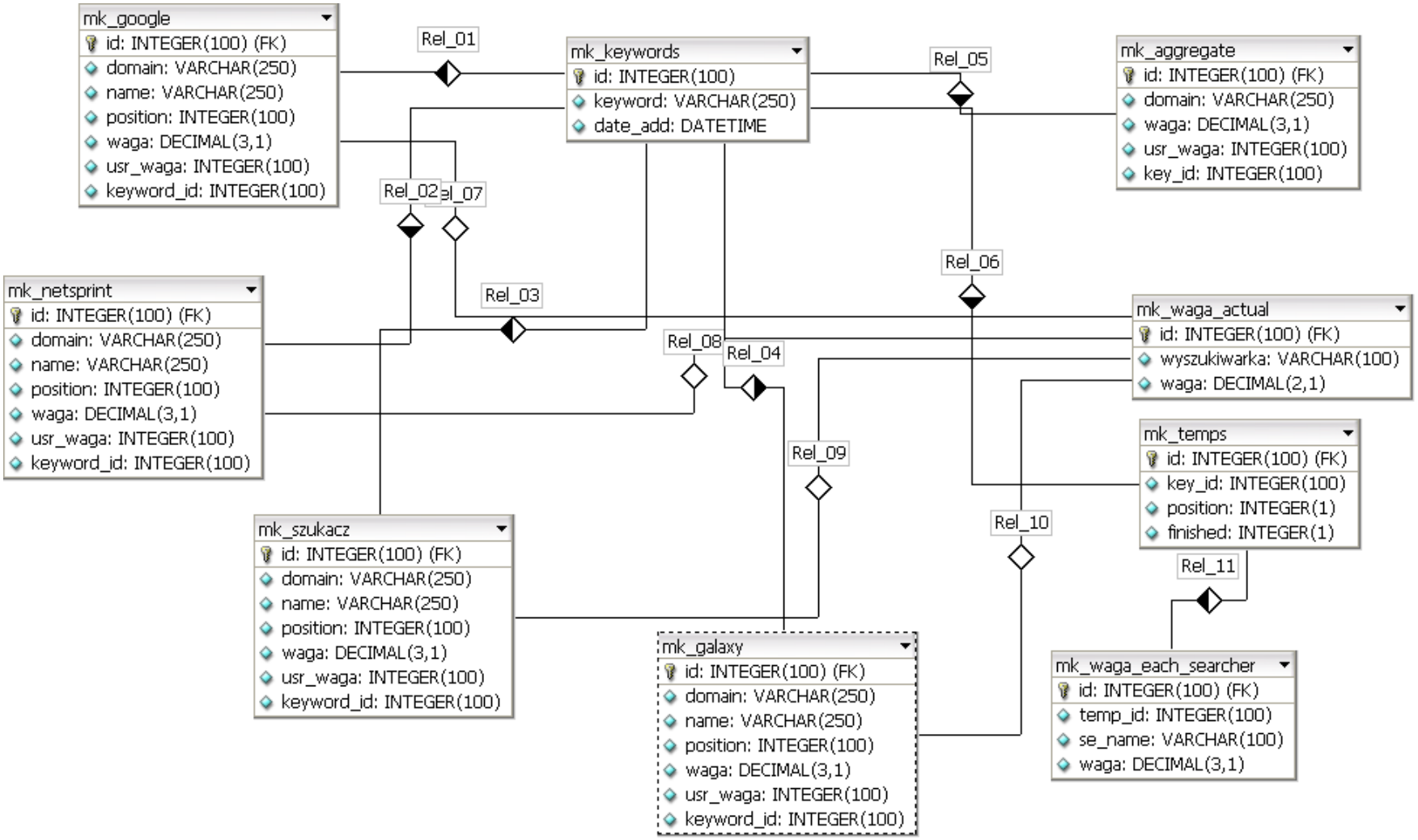
5.1.3 Biblioteki

1. jQuery.js

Jest to biblioteka dla języka JavaScript ułatwiająca komunikację pomiędzy Javascript i HTML. Posiada ona wiele funkcji zgodnych ze standardem Web 2.0 co sprawia, że staje się ona wielkim ułatwieniem przy tworzeniu dynamicznych stron WWW. Jednych z zalet, która została wykorzystana w systemie SEP, jest wsparcie dla modelu AJAX. JQuery dzięki wbudowanym funkcją obsługi modelu AJAX znacznie upraszcza i ułatwia wprowadzanie dynamicznych elementów do projektów webowych bez konieczności znajomości języka AJAX.

5.2 Struktura bazy danych

Baza danych w systemie SEP ogrywa ważną rolę w gromadzeniu informacji. Zaprojektowana baza składa się z 9 tabel typu MyISAM. Tabele zostały tak zaprojektowane aby umożliwić aplikacji szybkie i niezawodne działanie.



Rys. 2. Schemat struktury bazy danych dla systemu SEP.

Tab. 1. Struktura tabeli **mk_google**:

Nazwa pola	Typ pola	Opis
id	Int(100)	Identyfikator tabeli. Klucz podstawowy.
domain	Varchar(250)	Pełen Adres URL strony WWW.
name	Varchar(250)	Nazwa domeny strony WWW.
position	Int(100)	Pozycja strony w wyszukiwarce Google.pl.
waga	Decimal(3,1)	Ocena automatyczna dla danej strony.
usr_waga	Int(100)	Ocena użytkownika dla danej strony
keyword_id	Int(100)	Identyfikator słowa kluczowego.

Tabela **mk_google** służy do przechowywania danych skojarzonych z wyszukiwarka Google.pl. Gromadzone są tam dane uzyskane podczas procesu wyszukiwania oraz oceny uzyskiwane przez każdą ze znalezionych stron WWW.

Tab. 2. Struktura tabeli **mk_netsprint**:

Nazwa pola	Typ pola	Opis
id	Int(100)	Identyfikator tabeli. Klucz podstawowy.
domain	Varchar(250)	Pełen Adres URL strony WWW.
name	Varchar(250)	Nazwa domeny strony WWW.
position	Int(100)	Pozycja strony w wyszukiwarce Netsprint.pl.
waga	Decimal(3,1)	Ocena automatyczna dla danej strony.
usr_waga	Int(100)	Ocena użytkownika dla danej strony
keyword_id	Int(100)	Identyfikator słowa kluczowego.

Tabela **mk_netsprint** zbudowana jest identycznie jak tabela **mk_google** tyle że, jest ona skojarzona z wyszukiwarką netsprint.pl i przechowuje dane uzyskane i powiązane z tą wyszukiwarką.

Tab. 3. Struktura tabeli **mk_szukacz**:

Nazwa pola	Typ pola	Opis
id	Int(100)	Identyfikator tabeli. Klucz podstawowy.
domain	Varchar(250)	Pełen Adres URL strony WWW.
name	Varchar(250)	Nazwa domeny strony WWW.
position	Int(100)	Pozycja strony w wyszukiwarce Szukacz.pl.
waga	Decimal(3,1)	Ocena automatyczna dla danej strony.
usr_waga	Int(100)	Ocena użytkownika dla danej strony
keyword_id	Int(100)	Identyfikator słowa kluczowego.

Tabela **mk_szukacz** zbudowana jest identycznie jak tabela **mk_google** i **mk_netsprint** tyle że, jest ona skojarzona z wyszukiwarką szukacz.pl i przechowuje dane uzyskane i powiązane z tą wyszukiwarką.

Tab. 4. Struktura tabeli **mk_aggregate**:

Nazwa pola	Typ pola	Opis
id	Int(100)	Identyfikator tabeli. Klucz podstawowy.
domain	Varchar(250)	Nazwa domeny strony WWW.
waga	Decimal(3,1)	Sumaryczna ocena automatyczna dla danej strony.
usr_waga	Int(100)	Ocena użytkownika dla danej strony
key_id	Int (100)	Identyfikator słowa kluczowego.

Tabela **mk_aggregate** gromadzi dane połączone z trzema poprzednimi tabelami. Znajdują się w niej wszystkie nazwy domen jakie wystąpiły w wyszukiwarkach, ale bez powtórzeń, zsumowane oceny przyznawane automatycznie przez system oraz oceny od użytkowników.

Tab. 5. Struktura tabeli **mk_temps**:

Nazwa pola	Typ pola	Opis
id	Int(100)	Identyfikator tabeli. Klucz podstawowy.
key_id	Int (100)	Identyfikator słowa kluczowego.
position	Int (1)	Aktualna etap przebiegu analizy.
finished	Int(1)	Flaga zakończenia etapu.

Tabela **mk_temps** służy do kontrolowania dostępu użytkownika do poszczególnych obszarów procesu analizy podczas jej przebiegu. W pierwszej fazie kiedy użytkownik wyszukuje strony dla danego słowa kluczowego uaktywniany jest etap 1. Jeżeli system zakończy wyszukiwanie jest on potwierdzany w bazie odpowiednim wpisem w polu *finished* . W tym momencie użytkownik wkracza w etap 2, czyli ocenę wyszukanych stron WWW. Jednocześnie aby uchronić system przed ponowną próbą wpisania nowego słowa kluczowego i wyszukiwaniem go bez wcześniejszej oceny poprzednich rezultatów przez użytkownika, system SEP blokuje pole do wpisywania słów kluczowych. Nie zostanie ono odblokowane dopóki w bazie nie pojawi się flaga 1 w polu *finished* dla etapu 2. W tym momencie wiadomo że, użytkownik dokonał oceny rezultatów i może wyszukać nowe słowo. Uruchamiany jest również dostęp użytkownika do wykresu zależności ocen.

Tab. 6. Struktura tabeli **mk_waga_each_searcher**:

Nazwa pola	Typ pola	Opis
id	Int(100)	Identyfikator tabeli. Klucz podstawowy.
temp_id	Int (100)	Identyfikator z tabeli **mk_temps**.
se_name	Varchar (100)	Nazwa wyszukiwarki(google,

		netsprint, szukacz).
waga	Decimal(3,1)	Waga dla wyszukiwarki w danym etapie.

Tabela **mk_waga_each_searcher** służy do śledzenia wag wyszukiwarek w poszczególnych etapach analizy. Są to dane potrzebne do późniejszego tworzenia wykresu zmian wag na przestrzeni całego procesu analizy. Tabela zawiera identyfikator z tabeli **mk_temps**, co umożliwia skojarzenie wagi z danym etapem analizy oraz słowem kluczowym.

Tab. 7. Struktura tabeli **mk_waga_actual**:

Nazwa pola	Typ pola	Opis
id	Int(100)	Identyfikator tabeli. Klucz podstawowy.
wyszukiwarka	Varchar (100)	Nazwa wyszukiwarki.
waga	Decimal(3,1)	Aktualna waga.

Tabela **mk_waga_actual** jest tabelą „statyczną”. Znajdują się tam tylko 3 wpisy, każdy dotyczący innej wyszukiwarki. Tabela zawiera aktualną wagę dla każdej z wyszukiwarek. Wpisy do tej tabeli tworzone są w oparciu o sumę wag z poszczególnych etapów z tabeli **mk_waga_each_searcher**.

5.3 Struktura systemu

5.3.1 Opis poszczególnych elementów systemu

System SEP od strony użytkowej został zbudowany w oparciu o podstawowe aspekty ergonomi i funkcjonalności systemów informatycznych. Aby każdy nawet najmniej zaawansowany użytkownik mógł poprawnie korzystać z systemu, został on

zaopatrzony w 3 podstawowe elementy. Elementy te dzielą program na 3 osobne okna. W każdym z nich umieszczone są różne elementy pomagająca w przeprowadzaniu analizy wyszukiwarek internetowych.

Rys. 3. Podstawowe okno systemu SEP.

Po lewej stronie znajduje się panel o nazwie „wyszukiwarki” z wyszczególnionymi wyszukiwarkami.

Rys. 4. Panel lewy – wyszukiwarki.

W panelu znajdują się kolejno nazwy wyszukiwarek z których SEP korzysta, czyli Google.pl, Netsprint.pl, Szukacz.pl. Po kliknięciu na każdą z tych nazw rozwija się lista (rys.5). Lista ta zawiera wyniki wyszukiwania dla podanych słów kluczowych oraz oceny.

pozycja	domena	ocena auto.	ocena użyt.
1	wsse.krakow.pl	84.0	24
2	zik.com.ua	81.2	0
4	straz.bialystok.pl	75.6	30
5	strazacy.pl	72.8	3
6	eioba.pl	70.0	10
7	woda.org.pl	67.2	7
8	wszechnicagalicyjska.pl	64.4	30
9	katowice.uw.gov.pl	61.6	24
10	psseswidnica.pl	58.8	0
11	ww6.tvp.pl	56.0	0
12	przetargi.egospodarka.pl	53.2	0
13	nowysacz.pl	50.4	17
14	cksit.w.interia.pl	47.6	0
15	galabuda.pl	44.8	0
16	zegocina.pl	42.0	0
17	bip.ump.pl	39.2	0
18	pczkstaszow.info	36.4	29
19	erzeszow.pl	33.6	0
20	os.not.pl	30.8	0
21	flooding.ie	28.0	0
22	oki.krakow.rzgw.gov.pl	25.2	0
23	slowoludu.com.pl	22.4	0
24	silesiana.archiwa.net	19.6	8
25	ryglice.pl	16.8	21
26	msz.gov.pl	14.0	0
27	eko.org.pl	11.2	0
28	drinkingwaterquality.info	8.4	0
29	pwik.oswiecim.pl	5.6	0
30	wk.bip.um.klodzko.pl	2.8	0

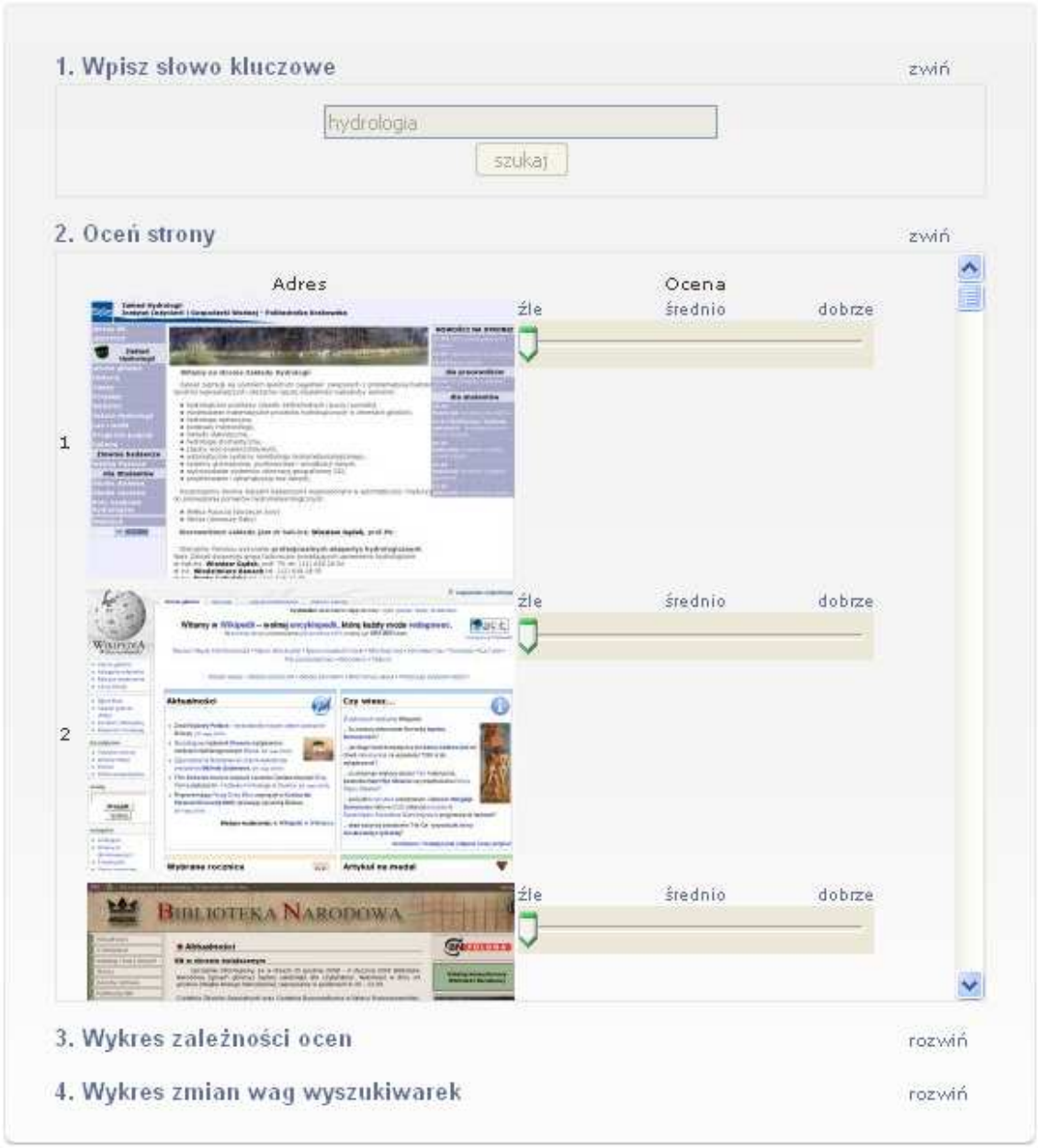
Rys. 5. Lista z wynikami wyszukiwarki Google.pl.

Jak widać na rysunku numer 5, lista zawiera wyniki dla aktualnie szukanej frazy. Na liście umieszczona jest także lista rozwijana z wszystkimi wyszukiwanymi do tej pory frazami. Po wybraniu z listy frazy, wyniki dla niej są wyświetlane poniżej.
Lista wyników zawiera:

- pozycję strony w wynikach danej wyszukiwarki
- nazwę domeny
- ocenę nadaną automatycznie przez system
- ocenę nadaną przez użytkownika

Każda domena oceniana jest przez system oraz przez użytkownika. Ocena użytkownika odbywa się w kolejnym kroku który jest szczegółowo opisany w rozdziale 5.4.3. System automatycznie ocenia każdą domenę. Na ocenę składa się iloczyn współczynnika dla pozycji na której znaleziona została domena oraz wagi. Waga domyślnie ma wartość 1.0. Jednak wraz z kolejnymi szukanymi słowami kluczowymi zmienia się. Może ona zwiększyć się lub zmaleć, w zależności od tego jak wyszukiwarka została oceniona przez użytkownika i jaki, w wyniku tej oceny, obliczony został współczynnik korelacji dla danej wyszukiwarki. Proces oceny użytkownika został szczegółowo opisany w rozdziale 5.4.3.
Dzięki panelowi „wyszukiwarki" użytkownik może w łatwy sposób sprawdzić wyniki dla każdej z wyszukiwarek i porównać oceny przyznane pierwotnie przez system a nadane przez samego siebie.

Drugim elementem jest centralny panel systemu (rys.6). Jest on największy ale i najważniejszy, gdyż to w nim widoczne są działania oraz operacje wykonywane przez system SEP. W panelu znajdują się cztery punkty które są jednocześnie tytułami kolejnych etapów analizy.



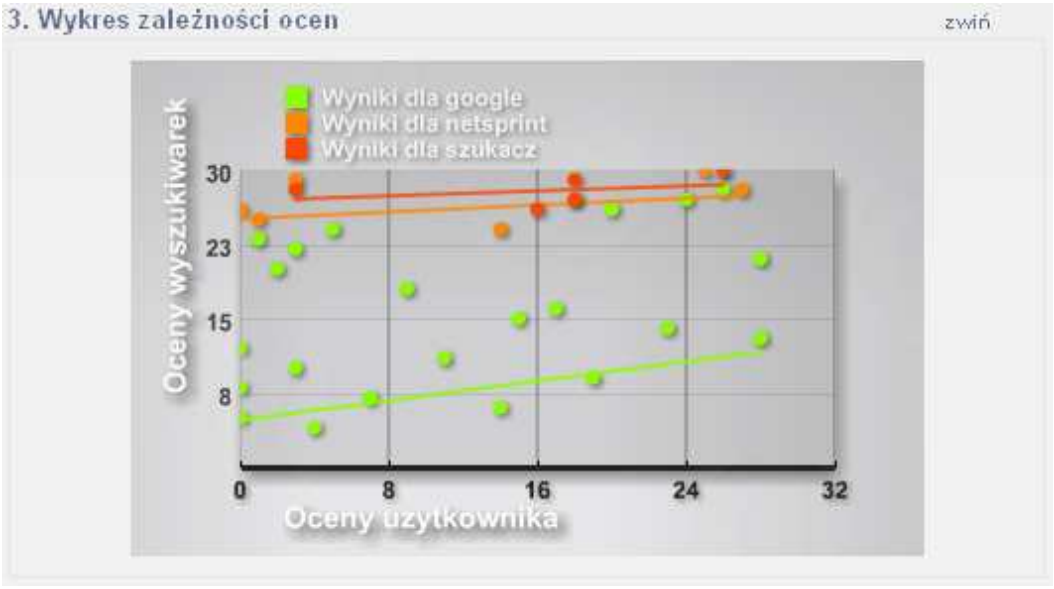
Rys. 6. Zawartość panelu środkowego.

Tytuły umieszczone w panelu opisują kolejne kroki procesu analizy. Ich kolejność nie jest przypadkowa, gdyż zawartość każdego z punktów uzależniona jest od operacji wykonanych na poprzednikach.

1. Pierwszym elementem jest „multiwyszukiwarka". Dzięki niej, podobnie jak w tradycyjnej wyszukiwarce (np. Google.pl) po wpisaniu interesującego nas hasła, wyszukiwane są strony WWW odpowiadające zadanym kryteriom. Element ten ma tytuł „Wpisz słowo kluczowe" i jest widoczny na rysunku numer 6.

2. Drugi element to okno z listą znalezionych przez „multiwyszukiwarkę” stron WWW ułożonych w kolejności od najwyższej do najniższej oceny nadawanej przez system. Lista składa się z liczby porządkowej, miniaturki strony oraz suwaka przy pomocy którego użytkownik nadaje oceny. Lista w zależności od ilości znalezionym stron jest długa lub krótka. Dlatego też lista wyposażona jest w suwak pionowy umożliwiający oglądanie całości wyników. Na końcu listy znajduje się przycisk „zapisz”, przy pomocy którego zapisywane są oceny dla poszczególnych stron w bazie danych. Element ten nosi nazwę „Oceń strony” i jest widoczny na rysunku numer 6.

3. Trzecim elementem jest wykres zależności ocen. Ujęte są w nim relacje pomiędzy ocenami nadanymi automatycznie przez system oraz ocenami nadanymi przez użytkownika.

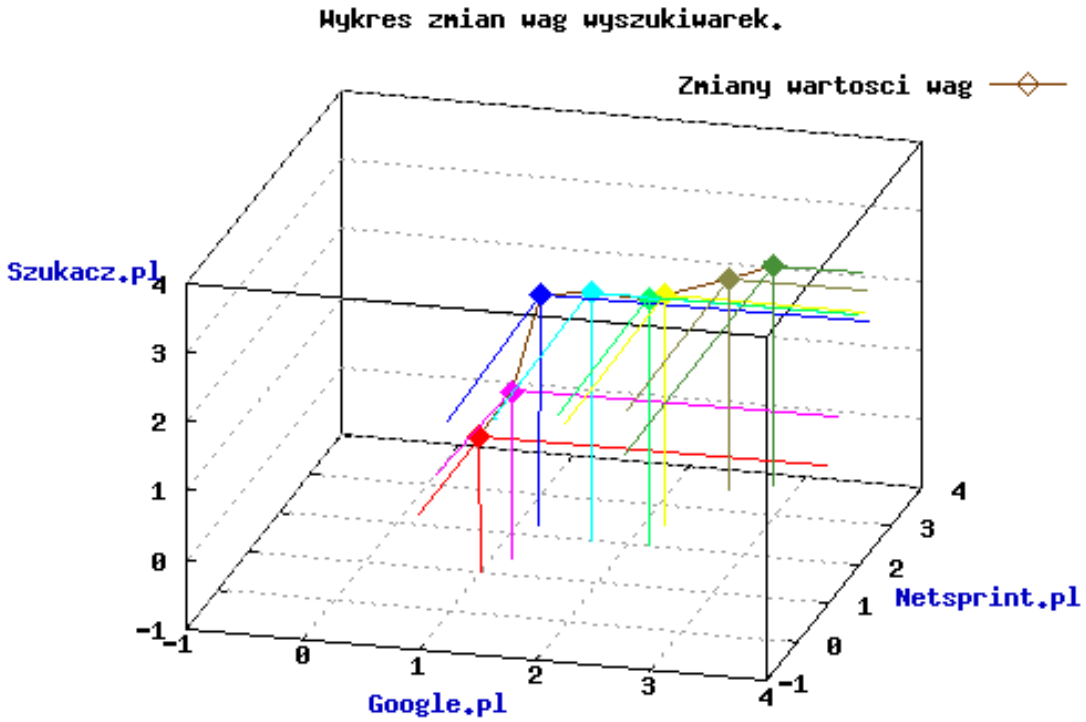


Rys. 7. Wykres zależności ocen automatycznych i ocen użytkownika dla wybranych wyszukiwarek.

Na wykresie widać jakie różnice są w ocenie nadanej automatycznie a nadanej przez użytkownika. Obie oceny składają się na wartości tworzących punkty na powyższym wykresie. Dzięki niemu widać rozbieżności w ocenie nadawanej przez system SEP automatycznie na podstawie rankingu danej wyszukiwarki oraz wagi, a oceną użytkownika.

Każda z wyszukiwarek oznaczona jest innym kolorem. Każdy punkt to para ocen, automatycznej i nadanej przez użytkownika, dla danej znalezionej przez wyszukiwarki strony WWW. Widoczne na wykresie linie pokazują trend jaki występuje w wynikach każdej z wyszukiwarek.

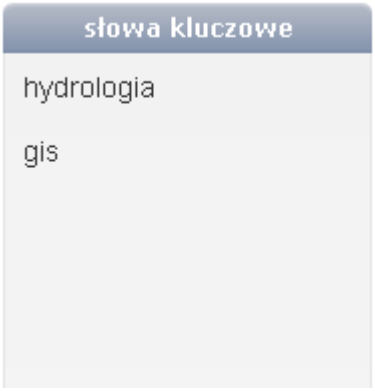
4. Czwartym elementem jest wykres obrazujący zmiany wag wyszukiwarek podczas analizy fraz. Wagi zmieniają się z każdą wyszukiwaną frazą. Są ona zależne od ocen nadanych automatycznie i ocen użytkownika. Wykres jest w trzech wymiarach, gdzie każda z osi należy do innej wyszukiwarki. Każdy punkt na wykresie to uporządkowany zbiór wartości wag badanych wyszukiwarek. Położenie tych punktów w przestrzeniu wykresu pozwala zobaczyć jak zmieniają się wagi wraz ze zmianami dokonywanymi w trakcie procesu analizy. Dzięki temu jesteśmy w stanie odpowiedzieć na pytanie która z wyszukiwarek dla zadanych przez nas fraz lepiej uwzględnia nasze oczekiwania co do rezultatów.



Rys. 8. Wykres zmian wag nadawanej każdej z wyszukiwarek w zależności od szukanej frazy.

Idealnym wynikiem byłaby zmiana wag z tendencją rosnącą. Oznaczałoby to iż, system SEP działa skutecznie i badane wyszukiwarki potrafią dostosować się do naszych oczekiwań.
Szczegółowy opis funkcji poszczególnych elementów okna, oraz ich działanie opisane jest w rozdziale 5.4.

Trzecim elementem od strony użytkowej jest prawy panel „słowa kluczowe”. W panelu tym znajdują sie kolejno ułożone słowa kluczowe/frazy wyszukiwane przez użytkownika.



Rys. 9. Panel prawy – słowa kluczowe.

Po kliknięciu na któreś ze słów rozwijana jest lista w której umieszczone są wszystkie znalezione przez wyszukiwarki strony WWW (rys. 10). Strony te są posortowane pod względem sum ocen ze wszystkich wyszukiwarek.

Rys. 10. Panel prawy z rozwiniętą listą.

Jak widać na rysunku numer 10, lista znalezionych przez wyszukiwarki stron WWW jest posortowana od tej z największą oceną. Jeżeli dana strona WWW powtórzyła się w rezultatach wyszukiwania w wyszukiwarkach, wtedy ich ocena jest sumowana. Ocena podawana w na liście jest oceną przyznawaną automatycznie przez system. Lista nie zawiera ocen nadawanych przez użytkowników. Oprócz ocen lista zawiera numer porządkowy oraz nazwę domeny.

Na każdym etapie korzystania z systemu SEP możliwy jest podgląd wyszukanych domen, co pomaga w jeszcze lepszym poznaniu badanych wyszukiwarek.

5.3.2 Model algorytmu „multiwyszukiwarki”

Zadania jakie wykonuje system SEP nakłada na niego obowiązek komunikacji z wyszukiwarkami internetowymi. To właśnie wyszukiwarki internetowe są obiektem badań. Aby móc dynamicznie sprawdzać ich działanie dla danych słów kluczowych, konieczne jest zapewnienie funkcjonalnej komunikacji z nimi.
Każda wyszukiwarka internetowa to na pierwszy rzut oka formularz składający się z większej lub mniejszej ilości pól. Formularz po wypełnieniu przez użytkownika jest wysyłany na serwer a tam przy pomocy odpowiednich skryptów interpretowany. Wysyłanie formularzy w języku PHP może odbywać się dwojaka, albo przy użyciu metody GET lub POST. Metody te różnią się od siebie tym iż, metoda POST nie umieszcza nazw oraz wartości poszczególnych pól formularza w adresie URL. Gdyby wyszukiwarki internetowe używały takiej metody niemożliwym byłoby dynamiczne generowanie zapytań do wyszukiwarek. Kolejnym problemem, odnoszącym się bardziej do potrzeb użytkownika jest to że, żaden użytkownik nie mógłby wysłać innemu użytkownikowi linka do znalezionych przez wyszukiwarki stron podając mu adres URL, gdyż nie zawierałby on żadnych wartości poza adresem strony wyszukiwarki. Jednak taki problem nie istnieje właśnie dzięki temu że, wyszukiwarki internetowe stosują metodę GET. Dzięki temu możliwe jest dynamiczne tworzenie zapytań.

Rys. 11. Adresy URL wyszukiwarek dla szukanego hasła „hydrologia”.

Jak można zauważyć na powyższym rysunku, każda wyszukiwarka ma inne nazwy parametrów potrzebnych w procesie wyszukiwania, oprócz parametru określającego szukaną frazę. Różna jest także ich liczba. Jedną z podstawowych cech łączących wszystkie badane wyszukiwarki jest to że, jako parametr zawsze posiadają słowo kluczowe lub frazę szukaną przez użytkownika w adresie URL. Fraza ta jest zakodowana tak aby była prawidłowo interpretowana przez przeglądarkę.

System SEP wykorzystuje własności wyszukiwarek i generuje dynamiczne adresy URL do każdej z badanych wyszukiwarek. Skrypt obsługujący taką komunikację napisany jest w języku PHP. Wykorzystywana jest do tego funkcja *fopen().* Jako parametry podawane są:

- strumień - pełen adres URL wraz z potrzebnymi dla danej wyszukiwarki parametrami.
- tryb „r" – określa sposób dostępu do strumienia, „r" wskazuje na tryb „tylko do odczytu".

Wykorzystując możliwości tej funkcji system uzyskuje informacje, na temat zawartości strony podanej w adresie URL. Ze względu na to że, adres URL generowany jest z parametrami takimi jak posiada normalny adres URL na stronie wyszukiwarki, system otrzymuje źródło strony wyszukiwarki. W pobranym w ten sposób źródle zapisane są wszystkie potrzebne dla systemu SEP informacje.

```
<link rel="prefetch" href="http://pl.wikipedia.org/wiki/Hydrologia">
<li class=g>
<h3 class=r>
<a href="http://pl.wikipedia.org/wiki/Hydrologia" class=l onmousedown="return
rwt(this,'','','res','1','AFQjCNECbyFjAjuPt_BTdcaC_iUfChNqKw','&amp;sig2=XY4Q_HY6wxJL8aGOVV7PPQ')">
<em>Hydrologia</em> - Wikipedia, wolna encyklopedia</a>
</h3>
<span class=m> <span dir=ltr>- 3 odwiedzin</span> <span dir=ltr>- 07-10-08</span></span>
<div class="s"><em>Hydrologia</em> (z gr. hydro, woda) - dział geografii fizycznej zajmujący
 się badaniem wody (pod każdą postacią) występującej w środowisku przyrodniczym. <b>...</b><br>
 <cite>pl.wikipedia.org/wiki/<b>Hydrologia</b> - 31k - </cite>
<span class=gl>
<a href=
"http://209.85.129.132/search?q=cache:QWfGJI97-v0J:pl.wikipedia.org/wiki/Hydrologia+hydrologia&amp;
hl=pl&amp;ct=clnk&amp;cd=1&amp;gl=pl" onmousedown="return
rwt(this,'','','clnk','1','AFQjCNEdKG86GTOyM7PJ-ymZ3kd_C_Ig9g','&amp;sig2=J6CUw56_5EWt5mglXEfbgw')">
Kopia</a>
 - <a href="/search?hl=pl&amp;lr=&amp;q=related:pl.wikipedia.org/wiki/Hydrologia">
 Podobne strony</a>
</span>
</div>
```

Rys. 12. Część plik źródłowego Google.pl podczas wyszukiwania stron WWW po wpisaniu hasła „hydrologia”.

Najważniejsza informacja zawarta w źródle strony wyszukiwarki, czyli ranking znalezionych stron dla każdej z wyszukiwarek jest inny. Na powyższym rysunku widać cześć kodu źródłowego wyszukiwarki Google.pl. Kod ten opisuje pierwszą pozycję w rankingu, który Google.pl utworzył po wpisaniu frazy „hydrologia”. Jak widać adres strony http://pl.wikipedia.org/wiki/Hydrologia jest zamknięty w znaczniki <a>. Przy pomocy wyrażeń regularnych system pobiera ze źródła strony tylko potrzebne dla niego wiadomości, czyli adres strony oraz jej pozycja w danej wyszukiwarce.

```
preg_match_all("#<h3 class=r>(.*)</h3>#U", $this->plik, $this->tablica);

foreach($this->tablica[1] as $si => $mon){
    preg_match("|href=\"http:\/\/(.*)\"|U", $mon, $this->tablica22[]);
}
```

Rys. 13. Fragment kodu odpowiedzialnego za wyłuskiwanie adresu URL ze źródła Google.pl.

Dla każdej z badanych wyszukiwarek wyrażenie regularne jest inne, ze względu na różnice w zawartości pliku źródłowego różnych wyszukiwarek.

Dane odczytane z plików źródłowych wyszukiwarek są zapisywane do bazy danych. Zapisywany jest pełny adres stron rankingu, adres domeny znalezionej strony oraz pozycja w rankingu danej wyszukiwarki.

Jednym z problemów który pojawia się podczas pobierania treść zawartej na stronach wyszukiwarek jest to, iż przy częstych zapytaniach do wyszukiwarek mogą one, ze względów bezpieczeństwa, zablokować adres IP z którego łączymy się z nimi. W taki sposób wyszukiwarki chronią się przed zbyt dużym i często złośliwym zwiększaniem liczby połączeń do strony.

5.4 Działanie systemu

System SEP to zestaw narzędzi, których zadaniem jest:

- umożliwienie użytkownikowi wyszukiwań dla dowolnej frazy/słowa kluczowego.
- pobranie (w tle) rezultatów wyszukiwania z trzech wyszukiwarek (Google.pl, Netsprint.pl, Szukacz.pl).
- uszeregowanie znalezionych stron pod względem pozycji w danej wyszukiwarce.
- odrzucenie powtarzających się adresów.
- nadanie automatycznie ocen dla każdej ze znalezionych stron, na podstawie przyporządkowanej przez wyszukiwarkę pozycji i wagi danej wyszukiwarki.
- możliwość oceny przez użytkownika znalezionych stron.
- obliczenie współczynnika korelacji na podstawie ocen przypisanych automatycznie i ocen nadanych przez użytkownika.
- stworzenie wykresu zależności pomiędzy ocenami przypisanymi automatycznie i ocenami nadanymi przez użytkownika
- stworzenie wykresu zmian wag dla każdej z wyszukiwarek

5.4.1 Schemat działania

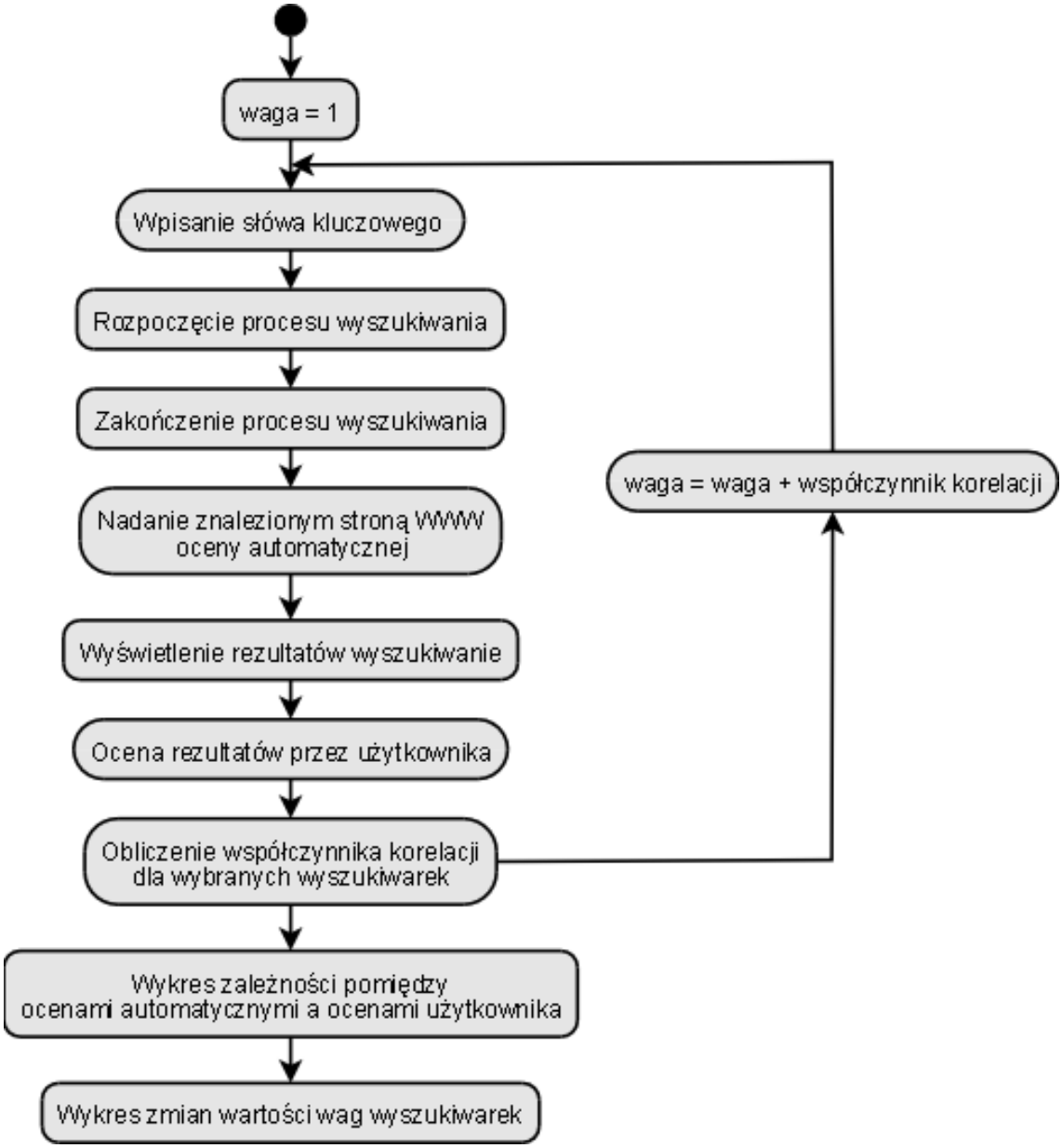
Dzięki SEP możliwe jest przeprowadzenie analizy systemów wyszukujących oraz zobrazowanie ich w sposób prosty i czytelny. Każda w opcji jakie posiada SEP zostały zaprogramowane tak aby użytkownik rozumiał ich działanie. Przejrzystość oraz prostota działania sprawiają, iż każdy kto będzie korzystał z systemu SEP z łatwością będzie mógł stworzyć analizę dla własnych potrzeb.

Każda wyszczególniona opcja nie działa w oderwaniu od pozostałych. Jeżeli każdy z elementów działałby osobno, a rezultaty jego działania nie miałyby wpływu na pozostałe, nie udałoby się uzyskać ciągłości w przetwarzaniu danych, a co za tym idzie wyników z poszczególnych etapów nie dałoby się ze sobą porównać.

W SEP należy wyróżnić 4 podstawowe etapy prowadzenia analizy:

1. Wpisanie słowa kluczowego.
2. Ocena znalezionych stron.
3. Wykres zależności ocen wyszukiwarek.
4. Zestawienie zmian wag wyszukiwarek.

Zależności pomiędzy etapami obrazuje schemat blokowy:



Rys. 14. Schemat blokowy działania systemu SEP.

Pierwsze dwa etapy czyli „Wpisanie słowa kluczowego” oraz „Ocena znalezionych stron” oraz ich wzajemna relacja to podstawowa cześć SEP bez zakończenia której niemożliwe jest przejście to kolejnych. Aby więc móc zobaczyć wykres zależności ocen dla wyszukiwarek należy najpierw wpisać słowo kluczowe a następnie ocenić rezultaty. Na podstawie danych od wyszukiwarek oraz danych prowadzonych przez użytkownika, tworzony jest wykres. Podobnie aby móc wpisać kolejną frazę do analizy, należy najpierw zakończyć proces analizy pierwszej frazy. Proces taki uważa się za

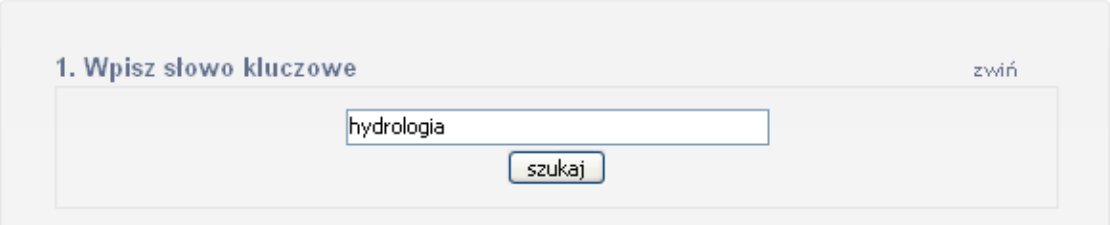
zakończony, w momencie zakończenia etapu drugiego czyli „Oceny znalezionych stron”. Zachowana ciągłość daje możliwość stworzenia ostatniego z etapów SEP czyli „Zestawienia zmian wartości wag wyszukiwarek”. Aby etap ten był możliwy do zrealizowania konieczne są co najmniej zakończone dwie analizy fraz.

5.4.2 Etap pierwszy - wyszukiwanie

System SEP ma za zadanie ułatwić odpowiedź na pytanie, która z wyszukiwarek dla danych słów kluczowych potrafi znaleźć jak najbardziej relewantne wyniki, a także która z nich potrafi uwzględniając poprzednie rezultaty poprawiać kolejne na podstawie określonych przez użytkownika oczekiwań. W systemie SEP badane są wyszukiwarki popularne, ogólnodostępne. Dzięki temu użytkownik ma szansę zbadać, czy dla słów kluczowych które podał owe wyszukiwarki są w stanie sprostać jego oczekiwaniom pod względem odszukanych stron WWW.
Aby takie badanie mogło być możliwe, SEP zaopatrzony jest w moduł multiwyszukiwarki. Moduł multiwyszukiwarki to cześć systemy która swoją funkcjonalności bardzo przypomina działanie zwyczajnej multiwyszukiwarki [3.4.6]. Jest jedną z podstawowych części systemu SEP, a to dlatego gdyż odpowiada za:

- odebranie i przetworzenie zapytania wysłanego od użytkownika
- stworzenie adresu URL z zapytaniem do wyszukiwarek dostosowanego do ich parametrów
- pobranie rezultatów z przeglądarek
- zapis wyników do bazy danych.

Moduł kolejno pobiera wartość jaką użytkownik podaje w polu formularza:



Rys. 15. Wygląd okna multiwyszukiwarki.

W pole użytkownik wprowadza interesującą go frazę/słowo kluczowe, które będzie przedmiotem analizy. Po kliknięciu przycisku szukaj, formularz zostaje wysłany a pole elementu 1 zostaję przysłonięte i pojawia się na nim informacja dla użytkownika że, proces wyszukiwania oraz przetwarzania danych właśnie się rozpoczął.

Rys. 16. Informacja okna wyszukiwarki o trwaniu procesu wyszukiwania

Zapewnia to, że użytkownik nie przerwie procesu pobierania informacji ze stron wyszukiwarek przed jego zakończeniem, poprzez wpisanie kolejnej frazy. Jest to także informacja na użytkownika, że proces wyszukiwania trwa.

W tym momencie system SEP rozpoczyna prace. Odebrana wartość pola którą podał użytkownik, jest w odpowiedni sposób przetwarzana tak, aby dostosować ją do każdej z badanych wyszukiwarek.

Kiedy zostanie już utworzony adres dostosowany do parametrów wyszukiwarek, rozpoczyna się proces wysyłania i odbierania wyników z wyszukiwarek. Algorytm odpowiedzialny na ten element systemu SEP został szczegółowo opisany w rozdziale 5.3.2. Rezultaty, czyli wyniki z wyszukiwarek są formatowane i sprawdzane tak, aby znalezione strony nie dublowały się.

System SEP oprócz pobierania wyników z wyszukiwarek, nadaje także automatycznie oceny każdej ze stron na liście wyników. Na ocenę składają się dwa czynniki.

1. Waga wyszukiwarki.
2. Pozycja strony WWW na liście rezultatów.

Waga wyszukiwarki początkowo (przed pierwszym wyszukiwaniem) jest domyślna dla każdej wyszukiwarki i wynosi 1.0. Waga zmienia się wraz z kolejnym wyszukiwanym słowem/frazą. W zależności od tego jak zostały ocenione przez użytkownika strony WWW otrzymane z danej wyszukiwarki, i obliczony na podstawie tego współczynnik korelacji, tak też zmienia się waga. Nowe waga, powstaje przez dodanie do siebie starej wagi i współczynniki korelacji obliczonego dla danej wyszukiwarki. Współczynnik korelacji szerzej opisany został w rozdziale 6.

Pozycja danej strony WWW jest oceniana punktowo z zakresu od 1 do 30 punktów. Gdzie strona znaleziona na 1 miejscu w danej wyszukiwarce dostaje maksymalną oceną. Lista branych pod uwagę stron jest zawężany do 30 pozycji, tak więc stronie na miejscu 30 przypada tylko 1 punkt.

Poprzez pomnożenie przez siebie dwóch czynników czyli wagi i punktów za pozycję w rankingu, otrzymujemy ocenę automatyczną – przyznawaną przez system SEP.

Po pracy wykonanej przez moduł multiwyszukiwarki następuje zapisanie danych do bazy danych. Model oraz szczegółowy opis struktury bazy znajduje się w rozdziale 5.2.

W bazie danych oprócz tego że, wyniki wyszukiwania są zapisywane dla każdej wyszukiwarki osobno to także sumaryczny wynik jest zapisywany. Wynik sumaryczny to zestawienie wszystkich znalezionych przez wyszukiwarki stron WWW posortowanych według oceny automatycznej. Jeżeli któraś z domen powtarzała się to ocena powstawała z sumy ocen w poszczególnych wyszukiwarkach.

Działanie multiwyszukiwarki oraz wszystkie procesy towarzyszące jej, napisane zostały przy pomocy technologii Ajax [4.3]. Dzięki temu procesy te są niewidoczne dla użytkownika, działają w tle, strona nie jest przeładowywana – nie następuje „zerwania” ciągłości w działaniu systemu.

5.4.3 Etap drugi – analiza wyników

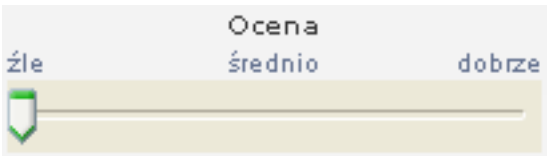
Opisane w rozdziale poprzednim dwa sposoby oceny znalezionych przez wyszukiwarki stron WWW, kształtują dalszą analizę sprawiając że, wzajemne zależności oceny automatycznej i oceny użytkownika stają się najważniejszym elementem systemu SEP. To dzięki tym różnicom możliwe jest przybliżenie się to stwierdzenia która wyszukiwarka jest lepsza. Analiza nie byłaby możliwa bez pobranych rezultatów z wyszukiwarek co zapewnia moduł multiwyszukiwarki. Jednak to ocena użytkownika rzutuje najbardziej na wyniki działania systemu.

Tytuł tego podrozdziału „Analiza wyników" jest dość ogólna, ze względu na to iż, analiza wyników przebiega tak naprawdę przez cały czas działania systemu. Jednak to w tym kroku drugim, użytkownik sam podejmuję decyzję która ze stron WWW spełnia najlepiej jego oczekiwania dla frazy którą wpisał. Użytkownik ma do dyspozycji miniaturkę rzeczywistej strony WWW.

Rys. 17. Miniaturki stron WWW.

Taka pomniejszona wersja oryginału zapewne nie może być przedmiotem zgłębionej analizy technicznej strony, ale może mieć wpływ na to czy użytkownik postanowi w ogóle przyjrzeć się tej stronie czy od razu ją odrzuci jak mało znacząca. Klikając na miniaturkę w przeglądarce otwiera się strona WWW w oryginalnym rozmiarze.

Narzędziem oceny strony które posiada użytkownik jest suwak.



Rys. 18. Suwak – narzędzie oceny dla użytkownika.

To rozwiązanie wpływa na łatwość przeprowadzania analizy i oceny stron. Ułatwia także wizualne zobrazowanie skali punktowej dla każdej z stron WWW.

Użytkownik ma do dyspozycji od 1 do 30 punktów. Punkty które nadaje użytkownik przy pomocy suwaka nie są widoczne, aby użytkownik mógł intuicyjnie ocenić czy dana strona jest „zła”, „średnia” lub „dobra”.
Im więcej punktów użytkownik przyzna danej stronie WWW, tym jego zdaniem ta strona jest dla niego bardziej wartościowa pod względem zawartej na niej treści.
Lista stron WWW sortowana jest domyślnie względem oceny automatycznej nadanej przez system SEP od najwyżej ocenionej.
Użytkownik zapisuje swoje oceny klikając w przycisk „Zapisz” znajdujący się na dole listy. Dane są zapisywane do bazy danych. Do tabel wyszukiwarek dopisywana jest ocena użytkownika. Również w tabeli sumarycznej dodawana jest ocena.
System posiadając ocenę użytkownika może rozpocząć proces obliczania współczynnika korelacji. Współczynnik ten liczony jest według wzory:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \times \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

gdzie,

x_i – ocena automatyczna dla stron i-tej,

y_i – ocena użytkownika dla stron i-tej,

n – ilość wszystkich stron podlegających ocenie

$\bar{x}$ - średnia arytmetyczna ze zbioru x,

$\bar{y}$ - średnia arytmetyczna ze zbioru y.

Jak zostało opisane to już w pracy, wagi wyszukiwarek początkowo są równe 1.0. Każde kolejne polecenie szukania wyników dla jakiejś frazy związane jest już z nowymi wagami dla każdej z wyszukiwarek, równymi sumie poprzedniej wagi i współczynnika

korelacji obliczonego według wzoru (1). Wagi są przechowywane w bazie i uaktualniane wraz ze zmianami na poszczególny etapach analizy.
System zmusza użytkownika do oceny stron dla podanej frazy. W momencie otwarcia się okna z listą stron, okno wyszukiwarki jest blokowane. Dopiero jak użytkownik oceni rezultaty działania wyszukiwarek i naciśnie przycisk „Zapisz", możliwe jest ponowne wyszukiwanie. Ma to na celu skłonienie użytkownika do przeprowadzenie oceny stron a nie jedynie wyszukania rezultatów dla zadanej frazy. Takie rozwiązanie daje pewność, iż system SEP będzie pracował w oparci o przemyślaną analizę użytkownika.
Po wykonaniu oceny stron przez użytkownika, następuje kolejny istotny proces, tworzenie wykresu zależności pomiędzy oceną automatyczną i nadaną przez użytkownika który wyświetlany jest w następnym, trzecim etapie.

5.4.4 Etap trzeci – wykres zależności ocen

Wykres zależności pomiędzy ocenami nadawanymi przez system a ocenami użytkownika ma na celu zobrazowanie różnic jakie wynikają z różnych aspektów badania stron WWW. Ocena automatyczna tworzona jest na podstawie pozycji danej strony w rankingu wyszukiwarki oraz wartości wagi wyszukiwarki. Ocena użytkownika natomiast to ocena spełnianych oczekiwań co do treści zamieszczonej na danej stronie WWW. Obie oceny składają się na szczegółowy obraz oceny wyszukiwarki pod kątem oczekiwań użytkownika. Im bardziej obie oceny są do siebie podobne tym lepiej oceniana jest wyszukiwarka.
Program do tworzenia wykresów nosi nazwę XML/SWF Charts. Został on zaprogramowany w technologii Flash. Umożliwia on tworzenie dowolnych wykresów na podstawie opracowanego pliku XML. Jest to darmowy program dostępny pod adresem www.maani.us. Zadaniem jego jest odwzorowanie macierzy w postaci pojedynczych punktów. Plik XML jest budowany dynamicznie na podstawie danych zapisanych w bazie danych.
Składa się on ze znaczników określających sposób formatowania wykresu, takich jak:

1. <chart_type></chart_type>

   Mówi o typie wykresu. W tym przypadku używany jest typ scatter.

2. <chart_border />

   Mówi o ramce wokół wykresu. Może zawierać kilka atrybutów (np.: color).

3. <chart_data></chart_data>

W znaczniku tego typu ujęte są dane widoczne na wykresie.

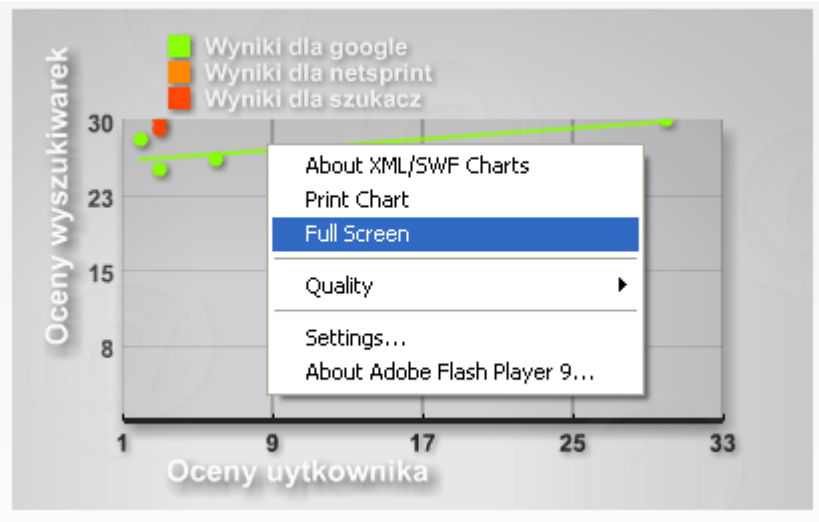
Dane ujmowane są w następujący sposób:

```
<row>
<string>Wyniki dla netsprint</string><number shadow='low' tooltip='15, 30.0'>15</number>
<number>30.0</number><number shadow='low' tooltip='11, 29.0'>11</number>
<number>29.0</number><number shadow='low' tooltip='10, 28.0'>10</number>
<number>28.0</number><number shadow='low' tooltip='16, 27.0'>16</number>
<number>27.0</number><number shadow='low' tooltip='11, 26.0'>11</number>
<number>26.0</number><number shadow='low' tooltip='20, 25.0'>20</number>
<number>25.0</number><number shadow='low' tooltip='24, 24.0'>24</number>
<number>24.0</number><number shadow='low' tooltip='13, 23.0'>13</number>
<number>23.0</number><number shadow='low' tooltip='5, 22.0'>5</number>
<number>22.0</number><number shadow='low' tooltip='10, 21.0'>10</number>
<number>21.0</number><number shadow='low' tooltip='22, 20.0'>22</number>
<number>20.0</number><number shadow='low' tooltip='10, 19.0'>10</number>
<number>19.0</number><number shadow='low' tooltip='7, 18.0'>7</number>
<number>18.0</number><number shadow='low' tooltip='5, 17.0'>5</number>
<number>17.0</number><number shadow='low' tooltip='2, 16.0'>2</number>
<number>16.0</number><number shadow='low' tooltip='3, 15.0'>3</number>
<number>15.0</number><number shadow='low' tooltip='2, 14.0'>2</number>
<number>14.0</number><number shadow='low' tooltip='2, 13.0'>2</number>
<number>13.0</number><number shadow='low' tooltip='3, 12.0'>3</number>
<number>12.0</number><number shadow='low' tooltip='4, 11.0'>4</number>
<number>11.0</number><number shadow='low' tooltip='3, 10.0'>3</number>
<number>10.0</number><number shadow='low' tooltip='11, 9.0'>11</number>
<number>9.0</number><number shadow='low' tooltip='6, 8.0'>6</number>
<number>8.0</number><number shadow='low' tooltip='3, 7.0'>3</number>
<number>7.0</number><number shadow='low' tooltip='10, 6.0'>10</number>
<number>6.0</number><number shadow='low' tooltip='3, 5.0'>3</number>
<number>5.0</number><number shadow='low' tooltip='3, 4.0'>3</number>
<number>4.0</number><number shadow='low' tooltip='3, 3.0'>3</number>
<number>3.0</number><number shadow='low' tooltip='2, 2.0'>2</number>
<number>2.0</number><number shadow='low' tooltip='0, 1.0'>0</number>
<number>1.0</number>
</row>
```

Rys. 19. Plik xml – zestaw znaczników służących do tworzenia wykresu XML/SWF Charts.

Plik flash importuje dane z pliku xml i tworzy na jego podstawie wykres.

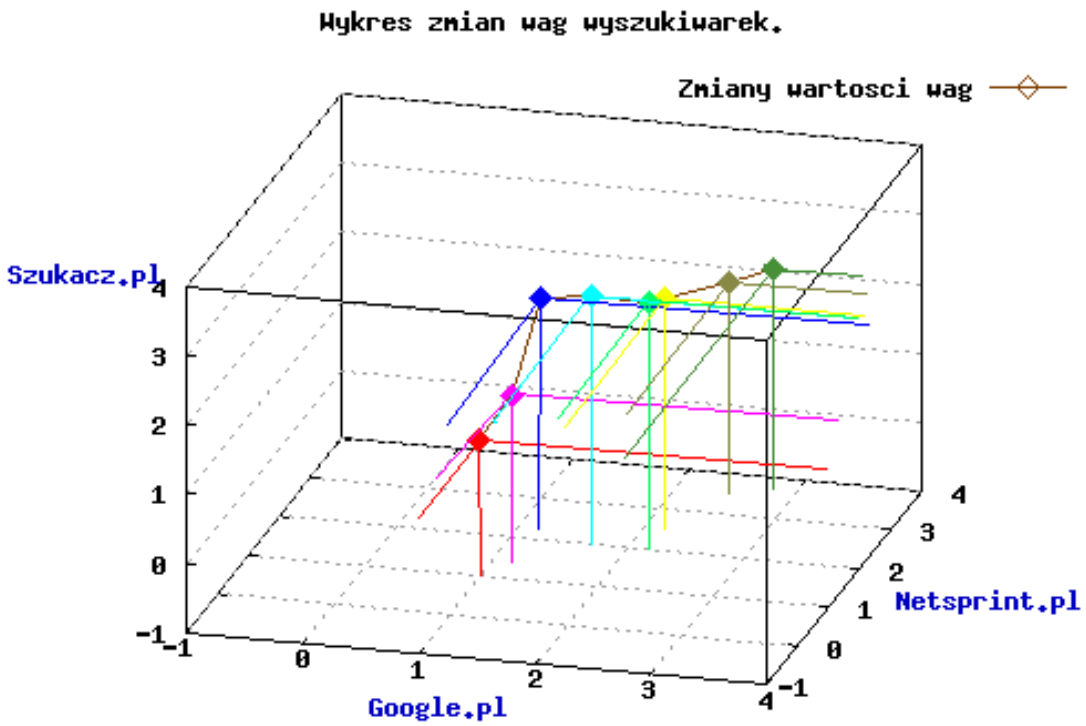
Na wykresie opisane są osie x i y, gdzie oś x reprezentuje oceny użytkownika natomiast oś y oceny automatyczne. Wykres można powiększyć po kliknięciu na niego prawym przyciskiem myszki i wybraniu opcji „Full Screen”.



Rys. 20. Opcja „Full Screen” na wykresie zależności ocen.

5.4.5 Etap czwarty – wykres zmian wag wyszukiwarek

Każda z wyszukiwarek po każdej zbadanej frazie jest oceniana przy pomocy wagi, na której wartość wpływ ma współczynnik korelacji. Współczynnik ten rośnie, maleje lub też pozostaje niezmieniony. Wykres widoczny poniżej do wykres punków składających się z trzech wartości: waga Google.pl, waga Netsprint.pl, waga Szukacz.pl.



Rys. 21. Wykres zmian wartości wag wyszukiwarek ujęty jako pojedyncze punkty.

Poszczególne wagi wyszukiwarek tworzą zbiory wartości które widoczne są na wykresie jako pojedynczy punkt. Poszczególne wartości wag wpływają na siebie. Rzutując punkty na każdą z osi, otrzymamy wartości wag dla poszczególnych wyszukiwarek.

Początkowym punktem jest punkt (1,1,1), gdyż takie początkowe wartości przybierają wagi wszystkich wyszukiwarek. Im współczynniki korelacji są wyższe po każdorazowej analizie frazy, tym zwiększają się wartości wag i punkty mają tendencję rosnącą. Szybki wzrost widoczny jest na wykresie w momencie bardzo dobrych rezultatów wszystkich wyszukiwarek. Brak zmian położenia punktów albo także zmniejszanie się ich wartości to znak, iż wyszukiwarki zostały źle ocenione.

Wykres tworzony jest automatycznie, każdorazowo po zapisaniu przez użytkownika ocen dla rezultatów stron. Dzięki przypisaniu każdej z osi wykresu do danej wyszukiwarki możliwe jest oglądanie wyników w formie obrazu 3D.

Narzędzie zastosowane do tworzenia wykresu nazywa się GNUPlot, które jest darmowym programem do tworzenia wykresów 2 i 3 wymiarowych.

6. Analiza działania systemu SEP dla określonych fraz

Kolejnym etapem pracy jest zbadanie jakie wyniki, pod względem potrzeb użytkownika dostarczają wybrane wyszukiwarki internetowe. Badanie to umożliwia system SEP. Testy dotyczą więc oprócz wyszukiwarek, także samego systemu SEP, gdyż to właśnie on udostępnia wiedzę na temat reakcji wyszukiwarek na zadane pytania, a także jest narzędziem które ma pomóc w zbadaniu możliwości wyszukiwarek do dostosowywania się do potrzeb użytkowników.

6.1 Obszary analizy

Analiza zostanie przeprowadzona przy założeniu, iż jest to tylko ułamek rzeczywistej naukowej analizy jaką należałoby przeprowadzić w celu jednoznacznego określenia odpowiedzi na zadane wcześniej pytania. Ilość badanych fraz jest więc mała, gdyż celem analizy nie jest zbadanie realnej umiejętności wyszukiwarek do dostosowywania się do potrzeb i oczekiwań użytkowników, ale jedynie stworzenie narzędzia które umożliwia taką analizę oraz zwrócenie uwagi na możliwości jakie niosłyby ze sobą takie umiejętności. Analiza jedynie kilku fraz z dziedziny inżynierii środowiska oraz ocena rezultatów prac wyszukiwarek tylko przez jednego użytkownika nie może być traktowana jako pogłębiona statystyczna praca na temat możliwości wyszukiwarek. Jednak narzędzie w postaci systemu SEP umożliwia takie próby i z pewnością mógłby posłużyć do bardziej rozbudowanych i zaawansowanych badań w tej kwestii.

Ze względu na to, iż analiza dogłębna wymaga sprawdzenia ogromnej ilości fraz na podstawie dopiero których można by wyciągnąć jednoznaczne wnioski, w pracy tej ujęte są zaledwie kilka z nich. Frazy pogrupowane są według przynależności do danej tematyki. Aby móc uzyskać informację na temat skuteczności wyszukiwarek, badane grupy składają się z frazy ogólnej oraz fraz bardziej szczegółowych z danego tematu. Badanie będzie przeprowadzane kolejno najpierw na frazie ogólnej a następnie na frazach szczegółowych danego obszaru inżynierii środowiska. Dzięki temu będziemy mieć możliwość sprawdzenia wyszukiwarek pod kątem poprawy swojego działania w

przypadku uszczegółowienia szukanego zagadnienia. Grupy fraz wybierane do analizy będą zazębiały się między sobą. Nie będą one przynależeć do zbytnio różniących się od siebie kategorii tematycznych. Ma to służyć skupieniu się na tylko wybranych obszarach nauki, w tym wypadku inżynierii środowiska.

Jeżeli wyniki nie będą satysfakcjonujące będzie to zapewne wina uproszczeń jakie zostały wprowadzona podczas analizy. Przede wszystkim będzie miało na to wpływ mała grupa wybranych fraz oraz krótki czas prowadzenia analizy. Jednak będzie to także wskazówka, która na pewno okaże się przydatna przy kolejnych próbach.

Przeprowadzona analiza będzie próbą odpowiedzenia na 2 podstawowe pytania:

1. Która wyszukiwarka najlepiej porządkuje wyniki dla tematyki inżynierii środowiska?
2. Czy optymalizacji wyników wyszukiwania w systemie SEP działa?

6.2 Analiza

Badanym obszarem naukowym w przeprowadzonej analizie jest szeroko pojęta inżynieria środowiska. Frazy wybrane do tego procesu są z różnych jej obszarów, jednak w pewien sposób są ze sobą połączone. Analiza wymaga stworzenie zestawy haseł, na których bazować będzie system SEP. Wybrane frazy to jedynie mały procent większej ilości, które należałoby sprawdzić aby uznać analizę wyszukiwarek za pełną, dogłębną i naukowo słuszną. Niemniej jednak cel, czyli próba odpowiedzi na pytania odnośnie systemów wyszukujących, poprzez subiektywne ich badanie za pomocą systemu SEP będzie spełnione.

6.2.1 Przebieg analizy

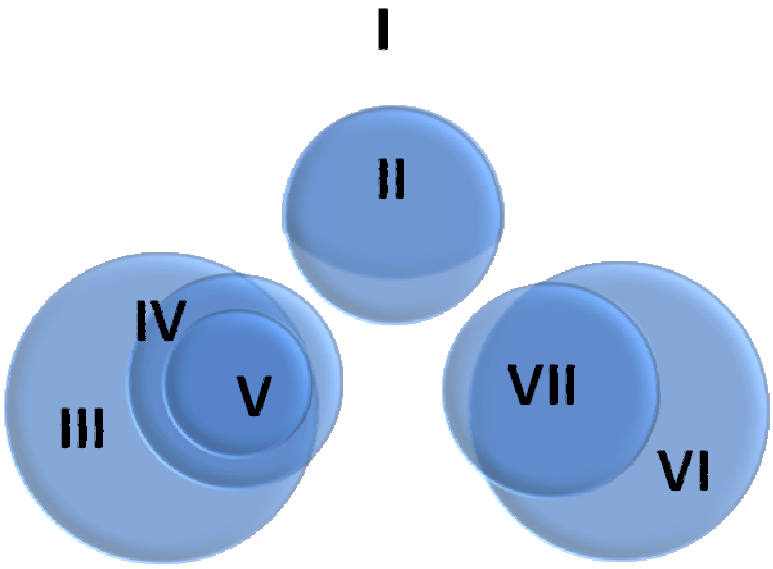
Analiza opierać się będzie o kilka zasad:

1. Badanie przeprowadzane jest poprzez kolejne wprowadzenie do systemu SEP siedmiu fraz/słów kluczowych.
2. Wyników wyszukiwania uzyskane przez badane wyszukiwarki oceniane są przez jednego użytkownika.
3. Frazy układane są w jedną wcześniej ustaloną kolejność badania.
4. Wyniki wyszukiwanie uzyskane przez badane wyszukiwarki są oceniane w sposób dokładny – przeglądana jest każda ze znalezionych stron.

Frazy które wzięły udział w badaniu (kolejno):

1. Pomiar objętości przepływu.
2. Czas przemieszczania fali wezbraniowej w górach.
3. Krążenie powietrza na Ziemi.
4. Od czego zależy siła wiatru.
5. Przyczyny powstawania tajfunów.
6. Jak samemu znaleźć wodę na działce.
7. Oczyszczanie studni po powodzi.

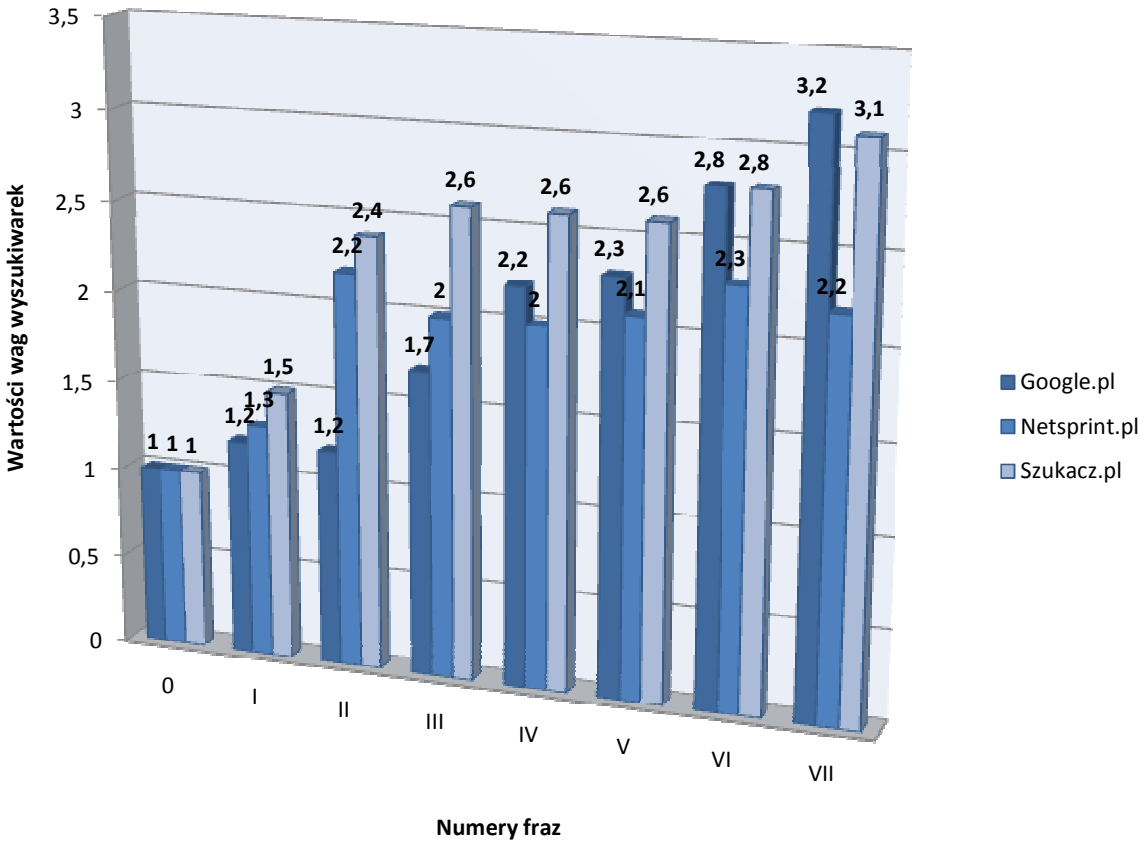
Frazy grupowane są w kategorie po dwie lub trzy. Każda z grup zawiera ogólne oraz uszczegółowione pojęcia z obrębu danej kategorii. I tak pierwsza grupa do frazy nr 1 i 2 druga grupa frazy nr 3, 4 i 5 oraz trzecia grupa to frazy nr 6 i 7.
Każda grupa do inna tematyka z zakresu inżynierii środowiska.



Rys. 22. Schematyczny podział fraz na grupy tematyczne: I - Pomiar objętości przepływu, II - Czas przemieszczania fali wezbraniowej w górach, III - Krążenie powietrza na Ziemi, IV - Od czego zależy siła wiatru, V - Przyczyny powstawania tajfunów, VI - Jak samemu znaleźć wodę na działce, VII - Oczyszczanie studni po powodzi.

Wybrane frazy oraz ich kolejność mają zasadniczy wpływ na sposób działania systemu SEP. Zostały one tak ułożone, aby dać możliwość weryfikacji wyszukiwarek pod kątem poprawy wyników wraz z uszczegółowieniem zadawanych fraz.

6.2.2 Analiza wyników



Rys. 23. Zmiany wartości wag badanych wyszukiwarek przy kolejnych badanych frazach.

Na powyższym wykresie naniesiony zostały zmiany wartości wag dla wyszukiwarek Google.pl, Szukacz.pl i Netsprint.pl w trakcie badania ich dla wybranych fraz. Frazy zostały na wykresie ponumerowane od I do VII, gdzie:

I. Pomiar objętości przepływu.
II. Czas przemieszczania fali wezbraniowej w górach.
III. Krążenie powietrza na Ziemi.
IV. Od czego zależy siła wiatru.
V. Przyczyny powstawania tajfunów.

VI. Jak samemu znaleźć wodę na działce.

VII. Oczyszczanie studni po powodzi.

Kolumna „0” to stan początkowy wszystkich wag dla wszystkich wyszukiwarek, czyli wartość 1.0.

Z wykresu można odczytać jak na podstawie wyników otrzymanych z wyszukiwarek dla podanej frazy oraz oceny ich przez użytkownika, zmienia się wartości wag tych wyszukiwarek. Począwszy od wartości początkowej równej 1.0 jego wartość systematycznie zmienia się w zależności jaka fraza była badana. Wyszukiwarki zachowywały się różnie w zależności badanej frazy. Raz zmiany wag był wysoki, rzędu 0,9, innym razem był równy 0 bądź nawet -0,2.

Należy przypomnieć, że wartości wag zależą od zmiany współczynnika korelacji wyszukiwarek które są badane. Współczynniki są obliczane na podstawie ocen przyznawanych automatycznie przez SEP oraz ocen użytkownika dla stron WWW znalezionych przez wyszukiwarki. Współczynnik korelacji jest z zakresu <-1;1>. Po kolejnych próbach, czyli po każdej badanej frazie współczynnik korelacji zmienia się, co oczywiście powoduję również zmianę wartości wagi wyszukiwarki.

W najlepszym wypadku, czyli kiedy współczynnik korelacji byłby równy 1, wartości poszczególnych wag rosłyby z każdą badaną frazą o 1. Jednak taki scenariusz jest mało prawdopodobny, gdyż oznaczałby, że wyszukiwarki są perfekcyjne w swych rezultatach, czyli kolejność wyświetlanych stron WWW idealnie pokrywałby się z oczekiwaniami użytkownika.

Jak pokazało badanie tylko przy niektórych z wybranych fraz, poszczególne wyszukiwarki bardzo dobrze reagowały na uszczegółowienie faz.

Jak widać, tylko Google.pl i Szukacz.pl nie miały podczas badania wartości ujemnych współczynnika korelacji, czyli zmiana wag tych wyszukiwarek była rosnąca. Po pierwszej zbadanej frazie współczynniki były dość niskie: Google.pl – 0,2, Netsprint.pl – 0,3, Szukacz.pl – 0,5. Taki wynik może świadczyć o tym, że wyszukiwarki słabo poradziły sobie z fraza ogólną nr I. Ciekawe wyniki dostarczyła próba druga z frazą nr II. Na podstawie wyników widać, że wyszukiwarki Netsprint.pl i Szukacz.pl w znakomity sposób spełniły oczekiwania użytkownika, gdyż współczynnik korelacji wyniósł aż 0,9

dla obu z nich co sprawiło, iż wagi tych wyszukiwarek wzrosły do wartości kolejno 2,2 i 2,4 . Google.pl natomiast poradziło sobie słabo.

Następny zestaw fraz które ułożone były od najbardziej ogólnej (nr III) do najbardziej szczegółowej (nr V), pokazał zupełnie odwrotną sytuację niż w poprzednich dwóch frazach. Badanie frazy nr III dało słabe wyniki, dla wyszukiwarki Netsprint.pl wręcz bardzo słabe (współczynniki korelacji równy -0,2).

Ujemny współczynnik korelacji w praktyce oznaczą, iż w dla badanej frazy oceny automatyczne znacząco nie zgodziły się z ocenami nadanymi przez użytkownika, a tym samym także z jego oczekiwaniami. Wynika z tego, że strony WWW uznane przez wyszukiwarkę jako mało znaczące i umieszczone na liście na dalszych pozycjach, w mniemaniu użytkownika były o wiele bardziej przydatne dla niego niż strony umieszczone na pierwszych pozycjach w rankingu wyszukiwarki.

Oczekiwany wzrost współczynnika przy kolejnych uściślonych w tematyce frazach nr VI i V nie nastąpił. Wyniki Google.pl początkowo na poziomie 0,5 spadły do wartości 0,1. Podobnie Szukacz.pl, którego wartość współczynnika stopniowo spadała aż do poziomu 0. Jedyną wyszukiwarką której wyniki poprawiały się, sądząc po wzroście współczynnika korelacji była Netsprint.pl. Wprawdzie wyniki są słabe ale jednak począwszy od frazy ogólnej i wartości współczynnika równej -0,2, poprzez kolejną frazę i wartość 0 aż do frazy nr V i wartości 0,1, wyniki poprawiały się.

W kolejnej badanie parze fraz nr VI i VII sytuacja powtórzyła się tak jak we wcześniejszej grupie fraz. Współczynniki jakie zostały obliczone po analizie frazy nr VI, czyli frazy ogólnej były przeciętne i kształtowały się na poziomie od 0,2 do 0,5. Najlepiej ocenione zostały wyniki otrzymane z wyszukiwarki Google.pl. Niestety uściślenie kolejnej frazy nie wpłynęło pozytywnie na wyszukiwarki. Zarówno Google.pl jak i Netsprint zmniejszyły trafność swoich wyszukiwań według użytkownika i ich współczynnik spadł. Google.pl z wartości 0,5 na 0,4, natomiast Netsprint.pl z 0,2 aż do wartości -0,1. Jedyną wyszukiwarką która w nieznaczny sposób poprawiła swoje wyniki jest Szukacz.pl, której współczynnik wzrósł z wartości 0,2 do 0,3.

Trudno na podstawie tylko kilku prób wyciągać wnioski co do działania wyszukiwarek oraz możliwości systemu SEO do personalizacji wyników wyszukiwania. Jednak biorąc

pod uwagę wyniki współczynników korelacji dla badanych fraz nie są one satysfakcjonujące.

Na słabe rezultaty największy wpływ ma zapewne dobór fraz oraz kolejność ich prowadzania do systemu. Frazy nie były tworzone z uwagą, iż w budowie powinny być odpowiednie, najbardziej korzystnych dla wyszukiwarek. Były to frazy dobrane do potrzeb i zawierające formę którą przeciętny użytkownik internetu stosuję podczas wyszukiwania.

Opierając się jednak o wykres (rys. 23), to wyszukiwarką która uzyskała najlepszy zmiany wag podczas przebiegu całego procesu analizy jest Google.pl. Jej współczynnik wyniósł 3,2. Niewiele gorszą wyszukiwarką okazał się Szukacz.pl którego wynik wyniósł 3,1. Najsłabiej podczas próby wypadł Netsprint.pl, którego wartość wagi wyniosła 2,2. Najlepszy wynik uzyskany przez Google.pl powinien wydawać się oczywisty, biorąc pod uwagę możliwości tej korporacji, jednak bardzo zbliżony wynik polskiej wyszukiwarki Szukacz.pl jest nieoczekiwany.

6.2.3 Wnioski

Na podstawie przeprowadzonej analizy trzech wyszukiwarek zbadanych siedmioma frazami z zakresu inżynierii środowiska, można stwierdzić, iż cel tajże pracy został osiągnięty.

Założone na początku dwa podstawowe cele, które badanie wyszukiwarek miało stwierdzić, czyli określenie która z nich działa najlepiej pod względem dostosowywania się do potrzeb użytkownika, oraz zbadania czy system SEP jest w pełni skuteczny w poprawianiu wyników otrzymanych od wyszukiwarek, zostały spełnione.

Okazało się, że adresy stron które zwracają wyszukiwarki są w bardzo małym stopniu zbieżne z oczekiwaniami użytkowników. Zazwyczaj strony, których użytkownik szuka są umieszczane na dalszych pozycjach rankingu stron WWW. Świadczą o tym bardzo niskie wartości współczynników korelacji podczas kolejnych prób oraz niskie wartości wag wyszukiwarek.

Z otrzymanych wyników można wnioskować, że wyszukiwarka Google.pl pomimo ogromnej przewagi finansowej, a co za tym idzie o wiele większymi możliwościami technologicznymi, nie zdobyła dużej przewagi nad o wiele „mniejszymi" konkurentami. Wynika z tego, że dla badanych fraz Google.pl nie sprostała oczekiwaniom użytkownika, podobnie jak pozostałe wyszukiwarki.

Ocena wyszukiwarek na podstawie, co prawda krótkiego badania jest jedna możliwa, a co za tym idzie pierwszy z celów pracy został osiągnięty.

Kolejny cel czyli, możliwości systemu SEP do optymalizacji wyników nie został do końca osiągnięty. Wagi po kolejnych próbach nie zwiększały się znacząco, niekiedy nawet malały. Wynikało to zapewne z doboru fraz oraz małej liczby prób, jednak wyniki osiągnięte przez system SEP nie są zadowalające. Warto jednak zaznaczyć, iż możliwości jakie daje system SEP są duże i w przypadku zwiększenia liczby prób oraz liczby użytkowników biorących w badaniu, rezultaty systemu byłyby na pewno lepsze.

7. Podsumowanie

Celem pracy było zaprojektowanie, stworzenie i przetestowanie systemu do badania trzech wyszukiwarek internetowych Google.pl, Netsprint.pl, Szukacz.pl oraz zbadanie możliwości systemu do personalizacji wyników wyszukiwania. System SEP został zaprojektowany w oparciu o dostępne technologie informatyczne. Dzięki wykorzystaniu języka PHP oraz AJAX system jest przyjazny użytkownikowi, czytelny w obsłudze oraz łatwy w migracji na różne platformy systemowe.

W pracy przybliżony został temat systemów wyszukujących oraz technologii związanej z rozwojem internetu. Na początku zapoznano się z rodzajami wyszukiwarek, budową systemów wyszukujących oraz zasadą działania tychże systemów. Dzięki teoretycznemu opisowi udało się łatwiej zrozumieć sens oraz kierunek naukowy niniejszej pracy.

Stworzony system o nazwie SEP umożliwia badanie wyszukiwarek internetowych. Dzięki zastosowanej technice obliczania współczynnika korelacji pomiędzy wynikami automatycznymi a oceną użytkownika, udało się opracować model badawczy który stwarza możliwość sterowania wynikami wyszukiwarek, a tym samym ocenę ich możliwości do dostosowywania się do potrzeb użytkownika.

Warto w tym miejscu nadmienić, iż w dniu 21 listopada 2008 czyli w momencie kończenia pisania tejże pracy, ukazał się artykuł na stronie: http://googlepolska.blogspot.com/SEPrch/label/wyszukiwanie opisujący najnowszy wynalazek firmy Google o nazwie WikiSearch, który jest narzędziem służącym do personalizacji wyników w wyszukiwarce tej firmy. Program wprawdzie stworzony zupełnie na inną skalę, jednak opierający swoją filozofię działania o możliwość dostosowywania wyników wyszukiwania do potrzeb użytkowników. Czyli jest to narzędzie którego prototyp został stworzony i opisany w niniejszej pracy. Warte także zaznaczenia, iż pomysł na napisanie tej praca powstał wcześniej niż system firmy Google oraz opiera swoje działanie na trzech wyszukiwarkach.

Zaimplementowany kompleksowy system SEP to oprócz możliwości przeprowadzania prób badań wyszukiwarek, dostarcza także wykresy obrazujące zmiany podczas poszczególnych prób oraz całego badania. Dodatkowo wykorzystanie bazy danych MySQL do gromadzenia danych sprawia, że za pomocą systemu można przeprowadzać dużo bardziej rozbudowane i dokładniejsze badania.

Test systemu oraz próba zbadania wyszukiwarek odbyła się na małej liczbie fraz. Jednak cel, czyli sprawdzenie czy system jest w stanie badać wyszukiwarki oraz umożliwiać użytkownikowi na wpływanie na rezultaty wyszukiwania został osiągnięty.

Bibliografia

1. Abiteboul Serge, Buneman Peter, Suciu Dan, „Dane w sieci WWW”, Wydawnictwo MIKOM, Warszawa 2001 r.
2. Bradenbaugh Jerry, „JavaScript Receptury”, Wydawnictwo HELION, Gliwice 2001 r.
3. Dubois Paul, „MySQL Podręcznik administratora”, Wydawnictwo HELION, Gliwice 2005 r.
4. Eichorn Joshua, „Ajax i JavaScript”, Wydawnictwo HELION, Gliwice 2007 r.
5. Glossbrenner Alfred and Emily, „Search Engines for the World Wide Web”, Peachpit Press, Berkeley 1998 r.
6. Kłopotek Mieczysław Alojzy, “Inteligentne wyszukiwarki internetowe”, Akademicka Oficyna Wydawnicza EXIT, Warszawa 2001 r.
7. Lecky-Thompson Ed, Eide-Goodman Heow, Nowicki Steven D., Cove Alec, “PHP5. Zaawansowane programowanie”, Wydawnictwo HELION, Gliwice 2005 r.
8. Ross Kenneth A., Wright Charles R.B., “Matematyka dyskretna”, Wydawnictwo Naukowe PWN, Warszawa 2003 r.

Netografia

1. Strona poświęcona wyszukiwarkom internetowym
   http://www.philb.com/webse.htm
2. Artykuł w magazynie PC World na temat wyszukiwarek internetowych
   http://www.pcworld.pl/artykuly/21318/Precz.z.bladzeniem.w.pajeczynie.html
3. Test wyszukiwarek internetowych magazynu Chip
   http://www2.chip.pl/arts/archiwum/n/articlear_18162.html
4. Strona poświęcona wyszukiwarkom
   http://www.spiders.pl/co_to_jest.php
5. Opis wyszukiwarki Google
   http://searchengineshowdown.com/features/google/review.html
6. Dokumentacja framework-a jQuery
   http://docs.jquery.com/Main_Page
7. Opis programu do tworzenia wykresów – Chart XML/SWF
   http://www.maani.us/xml_charts/
8. Oficjalny dokumentacja PHP
   http://php.net.pl/manual/pl/index.php

Spisy

Spis rysunków

Spis tabel