ZACHOWANIA UŻYTKOWNIKÓW WYSZUKIWARKI INTERNETOWEJ

Artur Strzelecki

Prace Naukowe / Uniwersytet Ekonomiczny w Katowicach

2010 I 122 - 131

Wprowadzenie

4 sierpnia 2006 r. zespół badawczy America OnLine, prowadzony przez Abdura Chowdhury, opublikował skompresowany plik tekstowy na jednej ze swoich stron, który zawierał 20 milionów słów kluczowych od ponad 650 tysięcy użytkowników, zbierany na przestrzeni 3 miesięcy. Dane były zbierane do celów badawczych. AOL zamknął publiczny dostęp do pliku po trzech dniach, ale stało się to, zanim zostały on skopiowany w inne miejsca Internetu.

Chociaż żaden z rekordów znajdujących się w pliku nie identyfikuje jawnie żadnego użytkownika, to jednak pewne słowa kluczowe zawierają informacje osobiste, które wprowadził do wyszukiwarki sam użytkownik, np. szukając informacji o samym sobie, o swoim adresie, numerze ubezpieczenia społecznego lub innych informacji osobistych. Ponieważ każdy użytkownik na tej liście identyfikowany jest przez unikalny klucz, moż-

liwa jest do odtworzenia historia wyszukiwania dla danego użytkownika. New York Times był w stanie odnaleźć indywidualne osoby z tego publicznego i anonimowego źródła na podstawie numerów telefonicznych abonentów. Powstało pytanie o etyczne korzystanie z tych danych do własnych badań.

Kolekcja danych zawiera informację o wpisywanych kwerendach przez użytkowników America OnLine (AOL) do zintegrowanej z AOL wyszukiwarki Google. Dane pochodzą od ponad 650 tysięcy użytkowników i zostały zebrane w okresie 3 miesięcy pomiędzy 1 marca 2006, a 31 maja 2006. Celem kolekcji jest dostarczenie prawdziwego zapisu danych na podstawie rzeczywistych użytkowników. Można te dane wykorzystać do personalizacji, formułowania kwerend lub innego rodzaju badań nad wyszukiwaniem.

Przegląd dziedziny

Zhang i Moffat [Zhan06] rozstrzygnęli pewne kwestie, które powstały na drodze interakcji użytkownika i wyszukiwarki na podstawie zapisanych logów i kliknięć. Ich obserwacje pochodzą z około 15 milionów zapytań zarejestrowanych w maju 2006 roku przez wyszukiwarkę MSN.

Pass, Chowdhury i Torgeson [Pass06] utworzyli wiele miar do opisania i ewaluacji efektywności i skuteczności dużych wyszukiwarek. Te miary, w ich pracy wizualne i werbalne, odkrywają obszary bogate złożoność i skalę. Odkryli sześć obszarów wyszukiwania: obszar kwerend, obszar sesji użytkownika, zachowanie użytkownika, wymagania operacyjne, obszar zawartości i demografię użytkowników.

Na podstawie udostępnionych danych przez AOL powstało również wiele dostępnych analiz w Internecie. Zbiór wszystkich takich obserwacji można znaleźć w angielskiej Wikipedii, w artykule „AOL search data scandal" [Wiki10].

Definicje

Odwołując się do terminów, konsekwentnie będziemy rozumieć pod nimi następujące definicje:

1. Kwerenda: Łańcuch wprowadzony przez użytkownika do wyszukiwarki jako prośba o informację.
2. Term: Pojedyncze słowo w kwerendzie, oddzielone białym znakiem. Termy mogą zawierać znaki alfanumeryczne, punktory lub inne symbole. Liczba termów w kwerendzie to długość kwerendy. Wyrażenia wielowyrazowe zawarte w cudzysłowiu również traktowane są jako kilka termów.
3. Sesja: Zbiór kwerend od jednego, konkretnego użytkownika uznanych (zwykle heurystycznie) za część pojedynczej interakcji z wyszukiwarką. Sesja może zawierać kwerendy, które dotyczą więcej niż jednej potrzeby informacyjnej.
4. Strona z wynikami: Uporządkowana lista rezultatów przedstawiona użytkownikowi w odpowiedzi na kwerendę. Strona z wynikami zwykle zawiera odnośniki do 10 rezultatów oraz zmienna liczbę odnośników sponsorowanych lub innego rodzaju.
5. Rezultat: Indywidualny adres URL na stronie z wynikami wraz z reprezentatywnym urywkiem tekstu wydobyty ze strony, zapewniający

dostęp do dokumentu zasugerowanego przez wyszukiwarkę jako odpowiedź na kwerendę.

6. Kliknięcie: Akcja użytkownika polegająca na kliknięciu w rezultat wyświetlony na stronie z wynikami, aby przejść do strony pod wskazanym adresem URL.

Bazując na tych definicjach zauważamy, że sesja zawiera jedną lub więcej kwerend, każda z nich składa się z jednego lub więcej termów. Każde wprowadzenie kwerendy powoduje powstanie strony z wynikami i jako rezultat przeglądania tej strony, użytkownik może wygenerować zero, jedno lub wiele kliknięć.

Zbiór danych

Zbiór AOL500k został udostępniony publicznie przez pracowników AmericaOnLine. W pliku znajdującym się przy zbiorze, widnieje informacja, że w przypadku używania tej kolekcji danych należy powołać się na artykuł „A Picture of Search" autorstwa Pass, Chowdhury i Torgeson [Pass06]. Ten zbiór danych zawiera ponad 21 milionów kwerend od użytkowników ze Stanów Zjednoczonych, zbieranych przez 3 miesiące i wprowadzanych do zintegrowanej z AOL wyszukiwarki Google. Wszystkie kwerendy znajdujące się w zbiorze posiadają znacznik czasowy, są przypisane do użytkownika oraz anonimowe. Każdy rekord zawiera (AnonID, Query, QueryTime, ItemRank, ClickURL):

- AnonID – numer ID anonimowego użytkownika,
- Query – kwerenda wysłana przez użytkownika,
- QueryTime – czas w którym, kwerenda została wysłana do wyszukiwarki,

- ItemRank – jeśli użytkownik kliknął w rezultat, zawiera pozycję klikniętego rezultatu na liście,
- ClickURL – jeśli użytkownik kliknął w rezultat, zawiera adres URL klikniętego rezultatu na liście.

Każdy rekord reprezentuje jedną z dwóch możliwych akcji:

1. Kwerenda za wynikami której użytkownik nie podążył i nie kliknął żadnego z rezultatów.
2. Kwerenda za wynikami której użytkownik podążył i kliknął jeden lub więcej rezultatów.

W pierwszym przypadku, tylko kwerenda, dane znajdują się tylko w trzech kolumnach (AnonID, Query i QueryTime). W drugim przypadku, kliknięcie rezultatu, dane znajdują się we wszystkich pięciu kolumnach. Gdy użytkownik kliknął więcej niż jeden rezultat na jednej stronie z wynikami, to na przykład będzie to reprezentowane dwoma rekordami danych dla dwóch kliknięć. Dodatkowo, jeśli użytkownik przeszedł na kolejną stronę z wynikami dla tego samej kwerendy, będzie to rekord z identycznie wprowadzoną kwerendą ale z późniejszym znacznikiem czasu.

Charakterystyka kolekcji

Przeglądając zbiór danych, szybko stało się jasne, że nie wszystkie kwerendy zostały wprowadzone przez interfejs AOL, zatem dane zawierają kwerendy, które zostały wysłane z zewnętrznych źródeł jak Web API, pasek narzędzi i inne oprogramowanie. Sesje przypisane do użytkownika o największej aktywności w ciągu tych trzech miesięcy zawierają łącznie 279430 kwerend. Oznacza to, że średnio jedna kwerenda od

tego użytkownika była wysyłana do wyszukiwarki co 28 sekund, co poświadcza, że działo się to w sposób zautomatyzowany. W trakcie największych sesji, kwerendy były wysyłane co sekundę lub kilka kwerend w trakcie tej samej sekundy. Tabela 1 zawiera podstawowe dane o badanej kolekcji danych.

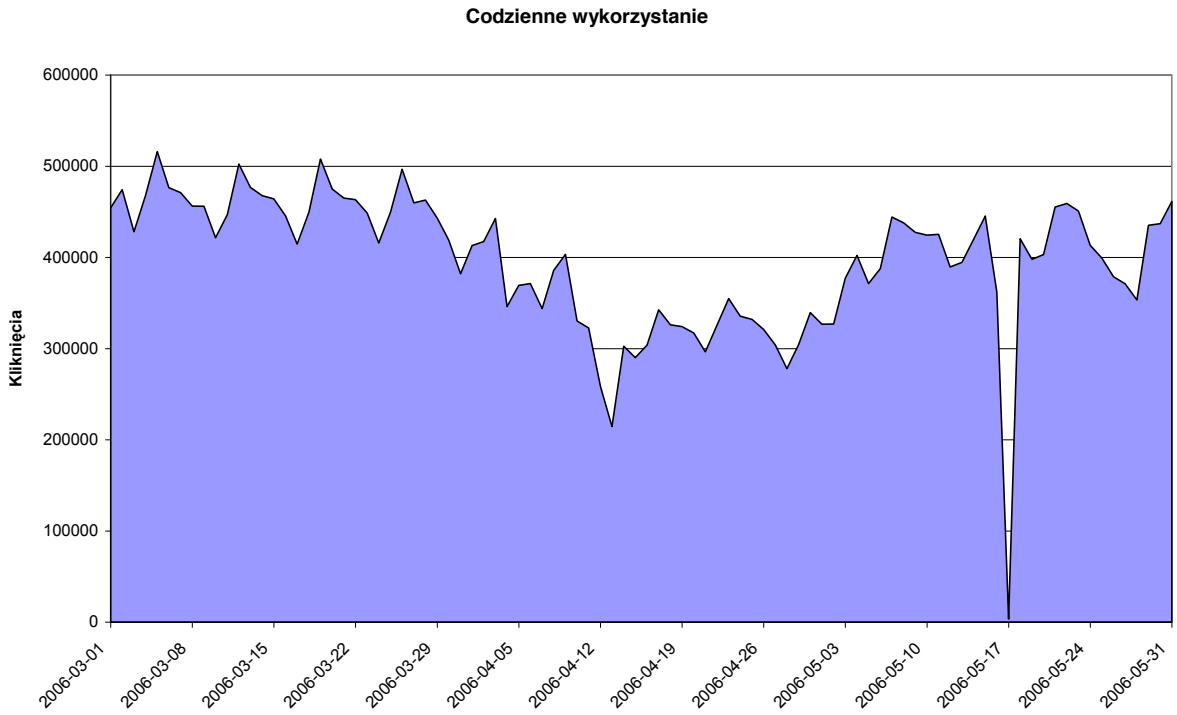
Tabela 1

Podstawowa charakterystyka kolekcji danych

Rekordy z danymi	36 389 567
Instancje nowej kwerendy	21 011 340
Przejścia na kolejną stroną z wynikami	7 887 022
Kliknięte rezultaty	19 442 629
Kwerendy bez kliknięcia	16 946 938
Unikalne kwerendy	10 154 742
Unikalni użytkownicy	657 426

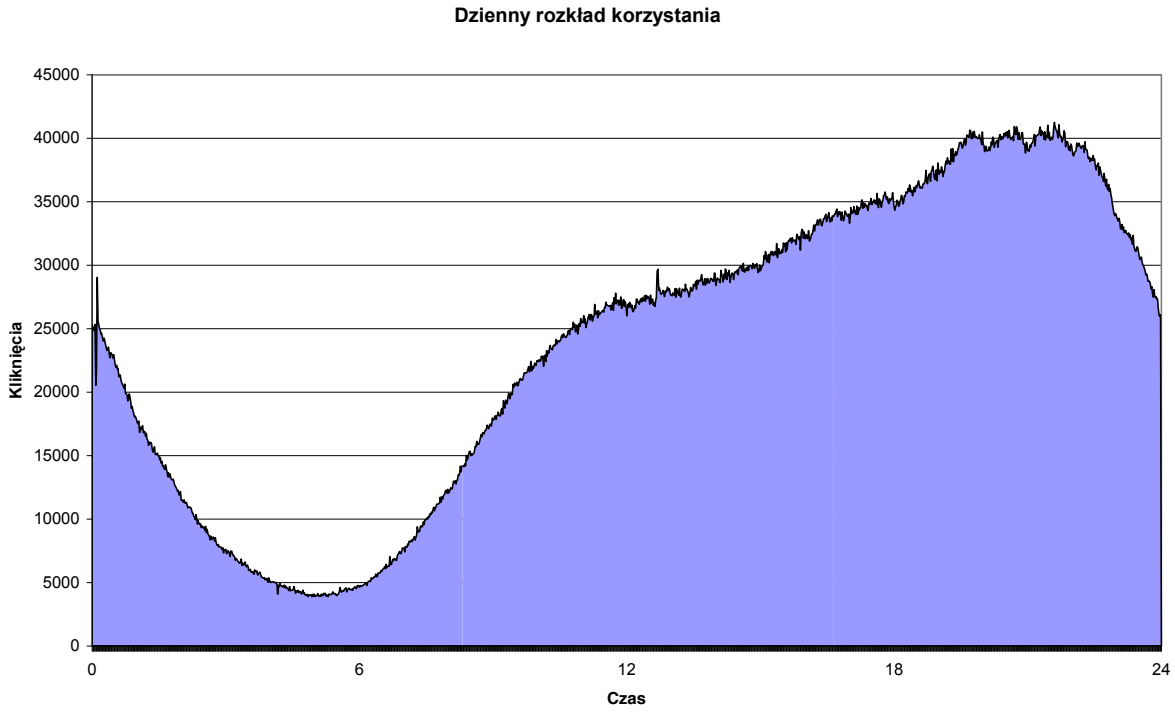
Źródło: Baza AOL500

Rysunek 1 pokazuje rozkład aktywności wysyłania kwerend w poszczególne dni każdego miesiąca. To oczywiste, że wolumen kwerend otrzymywanych przez wyszukiwarkę podąża za następującym wzorem, w dni robocze osiąga szczyt, natomiast w weekend silnie opada. Odzwierciedla to tygodniowy cykl pracy przeciętnego użytkownika, a nawet oddaje trend, zgodnie z którym, liczba wprowadzanych kwerend zmniejsza się wraz upływem czasu każdego tygodnia. Możliwe, że to wynik zmniejszania się indywidualnej produktywności. Sugeruje to także, iż zarówno wyszukiwanie w sieci stało się ważną i integralną częścią pracy w każdym standardowym biurze oraz że pracownicy wykorzystują zasoby pracodawcy (czas i łącze) do podejmowania swoich prywatnych poszukiwań w sieci. Dane z pewnością są niekompletne, ponieważ zawierają 100 razy mniej informacji w dniu 17 maj 2006.



Rys. 1. Codzienne korzystanie z wyszukiwarki,
Źródło: opracowanie własne, na podstawie danych AOL

Rysunek 2 ilustruje godzinowy rozkład wysyłania kwerend do wyszukiwarki. Szczegółowość skali jest na poziomie jednej sekundy a rozkład jednego dnia przedstawia uśrednione wartości dla całego okresu 92 dni w ciągu 3 miesięcy. Dokładna analiza danych ujawnia schemat przesyłania kwerend zgodnie z którym, aktywność korzystania z wyszukiwarki rozpoczyna się od wczesnego ranka, około godziny 4 rano czasu PST, w tym czasie w Nowym Jorku jest godzina 7 rano, osiąga swój szczyt około godziny 20 PST, gdy cały „kraj" jest w pracy, a następnie równomiernie opada aż do północy PST.



Rys. 2. Rozkład korzystania z wyszukiwarki w ciągu doby
Źródło: opracowanie własne, na podstawie danych AOL

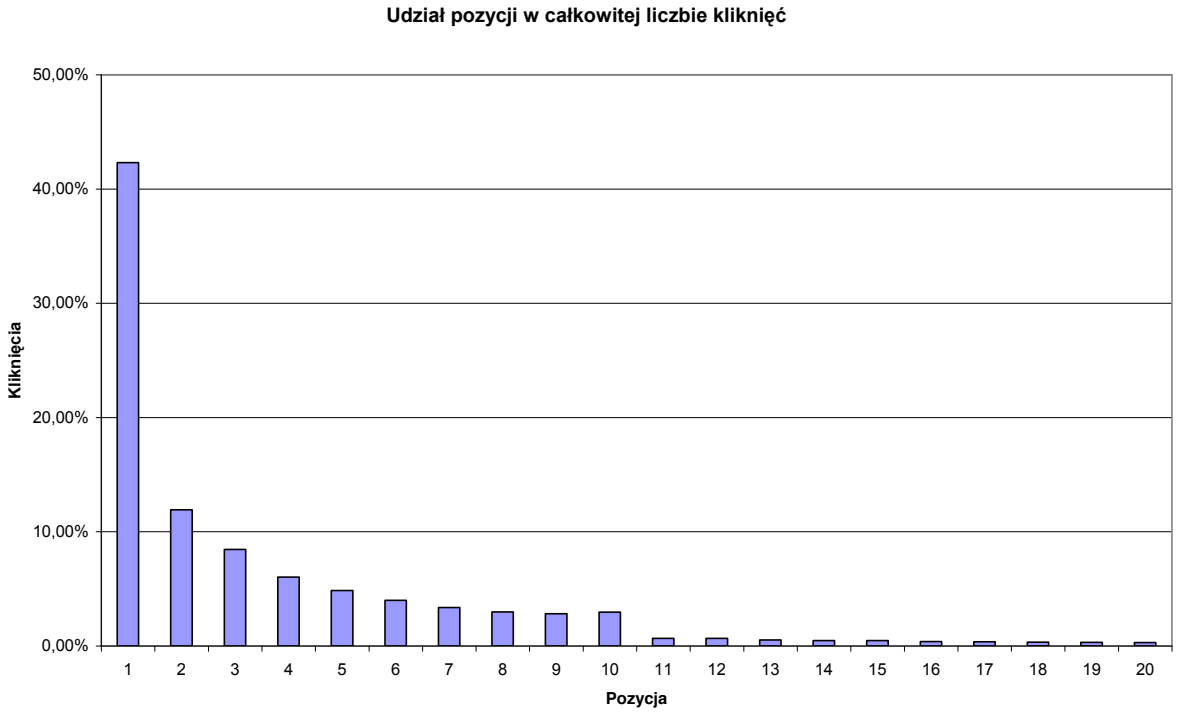
Kwerendy

Jedną z największych fascynacji przy analizie kwerend jest możliwość poznania, czego szukają użytkownicy. Pierwszą obserwację jaką można dokonać na podstawie tej próbki danych jest, iż kiedy pojawi się przed użytkownikiem okno dialogowe, często jest wykorzystywanie do kwerend nawigacyjnych. Dziesięć najpopularniejszych kwerend znajdujących się w danych Google, to były zapytania o inne popularne witryny internetowe, również o inne wyszukiwarki internetowe. Wiele z nich składa się z pełnego adresu URL.

Wpływ na marketing w wyszukiwarkach

Powszechnie wiadomo, iż najważniejsze i najczęściej odwiedzane są rezultaty na trzech pierwszych pozycjach w na stronie z rezultatami, te dane potwierdzają tę tezę. Analiza liczby klikniętych rezultatów pokazuje, iż rezultat na pozycji #1 zebrał 42,3% wszystkich zarejestrowanych kliknięć. Rezultat na pozycji #2 naliczył tylko 11,92% całkowitej liczby kliknięć – prawie 75% mniej kliknięć niż pierwszy rezultat. Z punktu widzenie widoczności w wynikach wyszukiwania, oznacza to, że rezultat na #1 pozycji zbiera czterokrotnie więcej wizyt użytkowników niż jego najbliższy rywal.

Rezultat na pozycji #3 osiąga 8,44% wszystkich kliknięć, prawie 30% mniej niż na pozycji #2 i ponad 80% mniej niż na pozycji #1. Przesuwając się w dół rezultatów na stronie z wynikami, każdy kolejny rezultat notuje mniejszą popularność od poprzedniego. Jednak należy zauważyć, iż rezultat na pozycji #10 zebrał niewiele więcej kliknięć niż rezultat #9. Najprawdopodobniej wynika to z reakcji użytkownika po przewinięciu strony z wynikami w dół, gdzie bardziej zauważa ostatni rezultat, zamiast przedostatni. Analizując drugą stronę z wynikami, należy zauważyć iż aktywność użytkowników dramatycznie spadła. Rezultat #11 wyniósł jedynie 0,66% ogólnej liczby kliknięć. Następujący wniosek z tego wynika, iż najbardziej liczy się obecność na pierwszej stronie z wynikami, pozostałe rezultaty są zauważane przez mniej niż 1% użytkowników.



Rys. 3. Klikalność kolejnych pozycji na stronie z wynikami

Źródło: opracowanie własne, na podstawie danych AOL

Użytkownicy wyszukiwarki najczęściej przechodzili na strony innych portali. Co ciekawe, AOL korzysta z silnika Google, jednak najczęściej klikanym przez użytkownika odnośnikiem było właśnie adres http://www.google.com. Tabela 2 przedstawia listę najczęściej klikanych adresów.

Tabela 2

Najczęściej klikane adresy

Adres	Kliknięcia
www.google.com	366623
www.myspace.com	167070
www.yahoo.com	161082
en.wikipedia.org	122539
www.amazaon.com	106119

www.imdb.com	98549
www.mapquest.com	91360
www.ebay.com	77947
mail.yahoo.com	53856
www.bankofamerica.com	48534

Źródło: Baza AOL500

Najwyraźniej użytkownicy traktują wyszukiwarkę jako nawigacyjny interfejs użytkownika i wpisują do niej prosto znane sobie adresy. Do stron wymienionych w tabeli 2 większość kwerend składała się fragmentu adresu url.

Zachowania użytkowników

Na podstawie analizy ścieżek wyszukiwania, czyli kolejnych przesyłanych kwerend oraz ich treści można wytypować 6 grup użytkowników korzystających z wyszukiwarki:

1. Oglądający pornografie – w bazie znajdują się miliony kwerend z zapytaniem o treści pornograficzne. Oglądających można nie tylko podzielić ze względu na to co szukają lecz także kiedy szukają. Niektórzy użytkownicy szukają pornografii przez cały dzień, inni tylko w sprecyzowanych przedziałach czasowych, które najczęściej rozpoczynają się około godziny 11 wieczorem.
2. Poszukujący innych osób – jedne osoby szukają innych osób. Większość danych pokazała, że szuka się osoby tylko raz, bez względu na porę i dzień, a następnie już się do niej nie wraca. Być może sprawdza się kandydatów do pracy lub szuka starych znajomych.
3. Kupujący – okazuję się, że kupujący najczęściej szukali okazji lub rabatów.

4. Obsesyjni – to osoby, które stale szukają tego samego.
5. Bezcelowi – wielu użytkowników nie ma obsesji na punkcie tego co szuka, po prostu stale używają wyszukiwarki i szukają nowych informacji, np. ciekawostek o filmach.
6. Nowi – ci użytkownicy dopiero dowiedzieli się jak włączyć komputer. Ci najczęściej wpisują w okno kwerendy adres url poszukiwanej strony lub np. nie używają spacji.

Inną z wartości jaką przedstawiają zebrane dane, jest zachowanie użytkowników jako konsumentów oraz jak konstruują kolejne kwerendy przy poszukiwaniu dóbr do zakupu. To bardzo przydatne, bo zazwyczaj można zetknąć się z danymi w postaci zagregowanej. Szukając produktów, np. laptopa, użytkownicy oczekują:

- znaleźć najlepszą ofertę,
- porównać marki,
- przeczytać recenzję,
- rozbudować dotychczasowy model,
- sprawdzić cenę,
- odwiedzić porównywarki cenowe.

Niektóre dane są niepokojące. Istnieje wiele rekordów, zawierających kwerendy użytkowników, którzy szukają informacji jak popełnić samobójstwo lub bardziej przerażające kwerendy, od osób które chcę popełnić zabójstwo.

Wnioski

Dane pochodzące z kwerend zapisywanych przez duże wyszukiwarki internetowe zawsze były dobrym materiałem do badań, dostarczając

istotnej wiedzy na temat interakcji między użytkownikami wyszukiwarki a nią samą. Wyniki badań na temat kwerend i klikanych rezultatów w wyszukiwarce mogą być stosowane w rożnych dziedzinach przetwarzania danych, włączając w to wkład w projektowanie interfejsu użytkownika, opracowanie fundamentalnego rankingu wyszukiwarki oraz usprawnienie metod buforowania i pobierania danych.

Literatura

[Pass06] Pass G., Chowdhury A., Torgeson C.: A Picture of Search, The First International Conference on Scalable Information Systems, Hong Kong, June, 2006.

[Zhang06] Zhang Y., Moffat A.: Some Observations on User Search Behavior, Proceedings of the 11th Australasian Document Computing Symposium, Brisbane, Australia 2006.

[Wiki10] http://en.wikipedia.org/wiki/AOL_search_data_scandal AOL search data skandal, dostęp 10.05.2010 r.

Informacje o autorze

Mgr Artur Strzelecki
Katedra Informatyki
Akademia Ekonomiczna
ul. Bogucicka 3
40-226 Katowice – Polska
Numer telefonu (fax) +48/32/2577277
e-mail: artur.strzelecki@ae.katowice.pl