Jak przeszukujemy sieć? „Długi ogon wyszukiwania"

Artur Strzelecki

Opracowanie[1] przedstawia analizę popularności witryn w wyszukiwarkach. Opiera się o dane pozostawione przez odwiedzających je użytkowników. Koncentruje się na użytkownikach, którzy odnaleźli te witryny w oparciu o mechanizmy wyszukujące w internecie. Prezentuje również pojęcie „długiego ogona" w wynikach wyszukiwarek oraz rezultaty przeprowadzonych badań dotyczących tego zagadnienia.

Każde zapytanie przesłane do wyszukiwarki pozwala w odpowiedzi uzyskać listę z wynikami wyszukiwania. Naturalne wyniki wyszukiwania, powstałe w oparciu o algorytm wyszukiwarki, który rankinguje i sortuje pozyskiwane rezultaty, prowadzą do adresów internetowych. Wyszukiwarki potrafią przemierzać usługi WWW, FTP i grupy dyskusyjne. Wybiórczo stosuje się opiniowanie redaktorów, których zadaniem jest sprawdzanie wysoko konkurencyjnych słów kluczowych w popularnych branżach. Inne dokumenty otrzymują natomiast ocenę rankingującą jako rezultat wyliczeń relewancji zapytania do znalezionych dokumentów.

Popularność wyszukiwarek internetowych jest silnie uzależniona od dokładności rezultatów, jakie przedstawiają. Im dokładniejsze wyniki, tym większą popularność zyskuje wyszukiwarka. Procedura ustalania rankingu jest fundamentalną charakterystyką każdej wyszukiwarki i ma ogromny wpływ na istnienie oraz popularność witryn w sieci. Wysoka pozycja na liście uzyskiwanych wyników dla słowa kluczowego związanego z witryną przynosi zazwyczaj znacznie więcej korzyści niż bardzo droga kampania reklamowa oparta o banery[2].

Ze względu na popularność poszukiwania wiadomości w sieci za pomocą wyszukiwarek pełnią one jedną z kluczowych ról w wirtualnym świecie. Przeprowadzone badania pokazują, że ponad połowa użytkowników odwiedzających po raz pierwszy witryny internetowe kierowana jest do nich prosto z wyszukiwarki. Wyszukiwarki odnotowują miesięcznie ponad 4,5 miliarda zapytań wprowadzanych przez użytkowników. Witryny konkurują ze sobą o to, do której z nich w wyniku wyszukiwania przejdzie użytkownik. Proste zapytanie skierowane do wyszukiwarki o dużych zasobach pozwala otrzymać tysiące, a nawet miliony odpowiedzi. Użytkownik sprawdza tylko kilka z nich, zwykle z pierwszej strony. 73% użytkowników wyszukiwarek nie analizuje wyników prezentowanych poza pierwszą stroną[3].

Optymalizowanie wyników w wyszukiwarkach

Celowe działania, które doprowadzają do znalezienia się na samym szczycie listy z wynikami wyszukiwania nazywamy pozycjonowaniem witryn internetowych. Pozycjonowanie witryn internetowych jest jednym z narzędzi realizacji strategii promocji przedsiębiorstwa w internecie. Polega ono na doborze działań zwiększających prawdopodobieństwo znalezienia się witryny wśród pierwszych wyników wyświetlanych przez wyszukiwarki internetowe. Wyniki tego działania dotyczą najbardziej popularnych słów kluczowych związanych z tematyką pozycjonowanego serwisu.

Według niektórych opracowań[4], czynniki, które potencjalnie wpływają na ranking wyszukiwarki in-

[1] Prezentowane w niniejszym artykule wyniki obserwacji witryn internetowych są częścią badań nad popularnością i widocznością witryn internetowych w wyszukiwarkach. Badanie jest oparte o dane pozostawione przez użytkowników odwiedzających witryny. W ramach tychże badań autor przeprowadził, poza analizą statystyczną, badania empiryczne, których częściowe wyniki zostały zaprezentowane podczas II Krajowej Konferencji Naukowej *Technologie Przetwarzania Danych* w dniach 24-26.09.2007.

[2] A. Khaki-Sedigh, M. Roudaki, *Identification of the dynamics of the Google ranking algorithm,* 13th IFAC Symposium On System Identification, Iran 2003.

[3] B.J. Jansen, A. Spink, *How are we searching the world wide web? A comparison of nine search engine transaction logs,* „Information Processing and Management" 2006, nr 42, s. 248–263.

[4] A. Bifet, C. Castillo, A. Chirita, I. Weber, *An analysis of factors used in search engine ranking*, 1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), Chiba 2005; S. Fortunato, M. Boguna, A. Flammini, F. Menczer, *How to make the top ten: approximating PageRank from In-degree,* Materiały z 14th International World Wide Web Conference, Edynburg 2006.

ternetowej dzielą się na dwie kategorie. Pierwsza to czynniki wewnętrzne (*query-factors*), które są zależne od treści w witrynie. Druga to czynniki zewnętrzne (*query-independent factors*), uzależnione od informacji pochodzących z zewnętrznych witryn, i posiadające odnośniki do reklamowanej witryny. Obie grupy czynników są trudne do wyliczenia, ponieważ wyszukiwarki internetowe nie ujawniają sposobu wykorzystania tych czynników do określenia pozycji w rankingu. Problem jest złożony z powodu następujących kwestii[5]:

- istnieje ponad 200 różnych czynników używanych m.in. przez wyszukiwarkę Google do wyliczenia rankingu strony;
- większość z nich jest nieznana i nie wiadomo dokładnie, jaki wpływ mają na wynik końcowy;
- waga każdego z użytych czynników do określenia wyników z pierwszej strony rezultatów wyszukiwania może być różna od wagi użytej dla pozostałych stron z rezultatami;
- różne zapytania mogą zostać obsłużone przez różne czynniki bądź różne wagi;
- Google posiada wielorakie centra z danymi rozmieszczone po całym globie, nie wszystkie są synchronizowane.

Zidentyfikowanie czynników zaangażowanych przez algorytm wyszukiwarki jest niezwykle trudne. W konsekwencji, zapotrzebowanie na wiedzę o tych czynnikach doprowadziło do powstania organizacji trudniących się tzw. *search engine optimization* (SEO) lub w szerszym ujęciu *search engine marketing* (SEM). Ich celem jest zwiększenie wartości rankingu w wynikach wyszukiwarek dla swoich klientów. Dzięki doświadczeniu i wielu testom są w stanie znacząco zwiększyć wynik wyszukiwania witryny. Trzeba pokreślić, że praca organizacji CEO ewidentnie opiera się częściowo na zgadywaniu, próbach i błędach. Pomimo tego, obroty na rynku SEO/SEM rosną z każdym rokiem.

Eksperyment i metodologia

Poniżej zaprezentowane zostały badania dotyczące analizy popularności witryn w wyszukiwarkach. Do badania wybrano 21 witryn, które były odwiedzane w okresie od 1 stycznia do 30 czerwca 2007 roku. Witryny są zróżnicowane pod względem budowy, treści i celu, w jakim zostały stworzone. Badanie opiera się na statystykach pobranych z systemu analitycznego Awstats, zainstalowanego na serwerze z usługą hostingową. Badane witryny zostały arbitralnie nazwane kolejnymi literami alfabetu, aby ułatwić czytelność i analizę porównawczą.

Pozyskanie danych do badania nie jest łatwe. Wszelkie informacje statystyczne właściciele witryn internetowych traktują jako poufne. Nie są one powszechnie dostępne. Dane dotyczące pierwszych siedmiu witryn autor uzyskał poprzez użyczony dostęp do systemu analitycznego, natomiast pozostałe informacje o 14 witrynach zostały znalezione w sieci. Odpowiednie zapytanie do wyszukiwarki wykazało wiele publicznie dostępnych statystyk niezabezpieczonych barierami technologicznymi. Po przejrzeniu około 100 statystyk autor wybrał 14 witryn z wiarygodnymi danymi do dalszego badania. Witryny przedstawiają różnorodną tematykę. Odrzucono strony z powiązaniami do stron pornograficznych oraz te, które nie miały żadnej treści.

Witryny zostały także ocenione pod kątem następujących kryteriów:

- liczba stron badanej witryny, zaindeksowanych przez wyszukiwarkę – niektóre witryny są większe niż inne, być może większe oznacza lepsze;
- wynik PageRank witryny w ocenie algorytmu Google opracowanego przez L. Page'a, S. Brina, R. Motwani, T. Winograda[6];
- liczba odnośników przychodzących, uzyskana operatorem 'link' w wyszukiwarce;
- wiek domeny witryny internetowej, spekuluje się, że starsze domeny są oceniane wyżej niż nowe w przypadku tej samej zawartości;
- obecność witryny w najważniejszych katalogach, edytowanych przez redaktorów, wpisy z katalogów Yahoo i DMOZ (*Open Directory Project*) zasilają wyniki w wyszukiwarkach Yahoo i Google. Witryny znajdujące się w tych katalogach przechodzą kontrolę wysokiej jakości. Strony tych katalogów są traktowane jako autorytety, które wyszukiwarki mogą wykorzystać jako jeden z czynników rankingu.

W badaniach uwzględniono następujące informacje:

1) Oglądalność, czyli:
   a) liczba unikalnych gości – użytkownicy, którzy odwiedzili stronę z niepowtarzalnym numerem IP,
   b) liczba wizyt – użytkownicy, którzy odwiedzili stronę, wraz z powracającymi do niej;
2) Źródła połączeń, w tym:
   a) liczba wizyt bezpośrednich lub z Ulubionych/Zakładek,
   b) liczba wizyt z wyszukiwarek,
   c) liczba wizyt z odnośników na innych stronach;
3) Popularność słów kluczowych i ich rozkład:
   a) poszukiwane słowa kluczowe (frazy),
   b) poszukiwane wyrazy.

[5] http://www.ankiety-online.pl

[6] Główny Urząd Statystyczny, *Warunki powstania i działania oraz perspektywy rozwojowe polskich przedsiębiorstw powstałych w latach 2001-2005* [online], Warszawa 2007, http://www.stat.gov.pl/cps/rde/xbcr/gus/PUBL_warunki_powstania_dzialania_przedsiebiorstw_2001-2005.pdf, [20.07.2007].

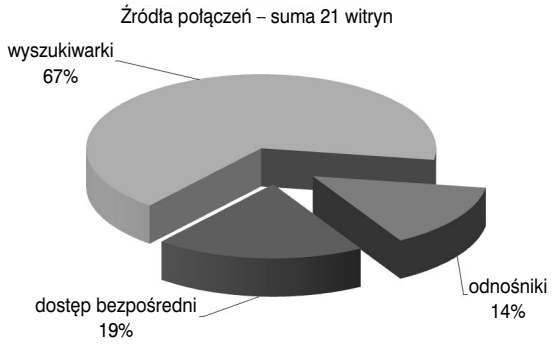
Oglądalność witryn

Oglądalność serwisu jest uzależniona od popularności witryny, częstotliwości, z jaką użytkownicy odnajdują daną stronę w internecie oraz samej budowy wewnętrznej witryny. Istnieje wiele innych stron ze znacznie większą oglądalnością niż te wzięte do analizy. Mimo to, już na uzyskanym poziomie wyświetleń, można wyprowadzić wiele cennych wniosków i porównań. Łączna suma wizyt w witrynach w badanym okresie wyniosła ponad 4,5 miliona. Liczba wizyt jest zawsze większa od liczby wizytujących użytkowników, różnica polega na zliczaniu wielokrotnych wizyt jednego, tego samego użytkownika na podstawie adresu IP oraz czasu odwiedzin.

Źródła połączeń

Do witryny internetowej można dotrzeć trzema sposobami. Pierwszy to wpisanie adresu URL do paska adresu przeglądarki internetowej, drugi to skorzystanie z umieszczonego w innym serwisie bezpośredniego odnośnika do witryny i trzeci, znalezienie adresu witryny w zasobach wyszukiwarki.

Rysunek 1. Źródła generowanych wizyt w witrynach



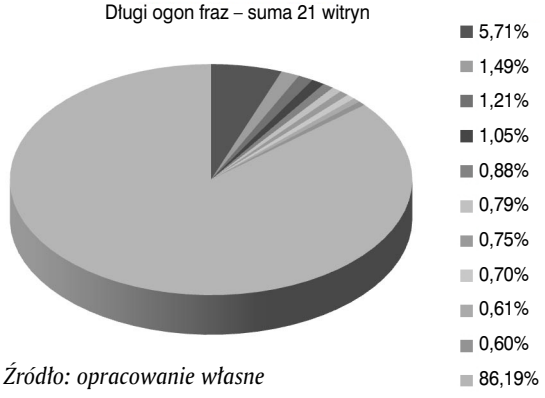
Źródło: opracowanie własne

Do wyszukiwarek prezentowanych na rysunku 1 należą: Goole, Live Serach, MSN, DMOZ, Yahoo, Alexa, AOL, Altavista i Seznam. W ich skład wchodzą również polskie wyszukiwarki OnetSzukaj, Szukacz i NetSprint. Średnia ważona rozkładu źródeł połączeń wskazuje na kluczową rolę wyszukiwarek w docieraniu użytkowników do witryn internetowych. Ponad 2/3 użytkowników korzysta z wyszukiwarek żeby odnaleźć pożądaną stronę. Dostęp bezpośredni to wejścia na witrynę przez użytkowników, którzy adres do strony mają zapisany w zakładce ulubione swojej przeglądarki internetowej, zapamiętany adres w historii przeglądanych witryn lub znają adres na pamięć i wpisują go w pasek adresu przeglądarki. Dostęp przez odnośniki to przejście z jednej witryny na drugą z wykorzystaniem odsyłacza hipertekstowego.

Frazy kluczowe

Witryny budowane w technologii tekstowej, wypełnione wysokiej jakości pożądanymi treściami, mają przy wyszukiwaniu przewagę nad witrynami, które nie mają takiej zawartości. Liczba utworzonych słów kluczowych, czyli fraz, dzięki którym witryna została znaleziona w wyszukiwarkach we wszystkich przypadkach jest średnio 4–5 razy większa od całkowitej liczby słów znalezionych na tej stronie. Użytkownicy wykorzystują różne kombinacje słów kluczowych. Liczba możliwych kombinacji, jakie jeszcze mogą powstać, rośnie w postępie geometrycznym.

Rysunek 2. Procentowy udział poszczególnych fraz



Źródło: opracowanie własne

Wyniki na rysunku 2 prezentują procentowy udział wizyt wygenerowanych przez pierwsze dziesięć najpopularniejszych fraz ze słów kluczowych do wizyt wygenerowanych przez pozostałe frazy. W badanej próbce dwie witryny wykazywały się anomalią. Pierwsze słowo kluczowe wygenerowało około 50% (O i P) ruchu z wyszukiwarek dla obu. Średnia ważona pokazuje, że pierwsze dziesięć słów kluczowych daje niecałe 14% ruchu w witrynach, natomiast zdecydowana większość, ponad 86% ruchu z wyszukiwarek pochodzi ze słów kluczowych spoza pierwszej dziesiątki.

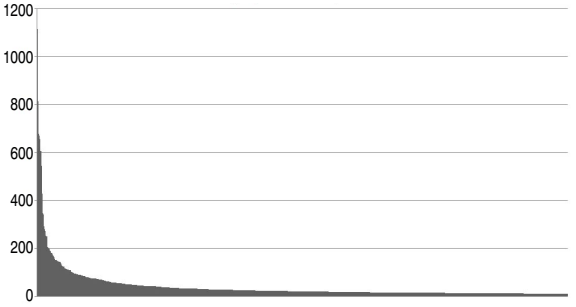
„Długi ogon wyszukiwania”

Witrynę internetową można opisać za pomocą kilku słów kluczowych, tych najważniejszych, związanych z jej tematyką. Wykres pokazuje, że kilka czy kilkanaście głównych słów kluczowych generuje niewielki udział w oglądalności witryny. Nie ma tu zastosowania zasada Pareto. Zasada Pareto, znana również pod nazwą zasady 80-20, utrzymuje że 80% efektów pochodzi z 20% nakładów. Przedstawione wyniki pokazują, że wszystkie zaplanowane i celowe działania stworzenia właściwego opisu za pomocą najbardziej trafnych słów kluczowych generują tylko 14% ruchu z wyszukiwarek.

Zaistniałą sytuację nazwano „długim ogonem wyszukiwania”. „Długi ogon” to termin, którego nazwa pochodzi od wykresu w układzie współrzędnych XY, na którym po początkowym gwałtownym spadku z wysokiego poziomu, w dalszej części następuje wydłużone, bardzo powolne zbliżanie się do zera, co można zaobserwować na rysunku 3. Przedstawiony wykres obrazuje częstotliwość fraz doprowadzających do witryny T. Autorem tej koncepcji jest Chris Anderson, redaktor naczelny magazynu „Wired”.

W pierwotnej wersji „długi ogon" odnosi się do modelu biznesu realizowanego w internecie, choć opisuje zjawisko od dawna znane w statystyce. Zgodnie z tym modelem, niektóre przedsiębiorstwa internetowe, jak choćby księgarnia Amazon.com, połowę swoich obrotów uzyskują za sprawą najpopularniejszych i intensywnie reklamowanych produktów. Jednak druga, o wiele bardziej zróżnicowana część rynku, składająca się z produktów niszowych, jako całość generuje nie mniejsze wpływy. Długi ogon to miejsce dla relatywnie mało znanych produktów, które w globalnej gospodarce sieciowej mogą znaleźć jednak znaczną liczbę amatorów[7].

Rysunek 3. Długi ogon wyszukiwania, przykład witryny T



Źródło: opracowanie własne

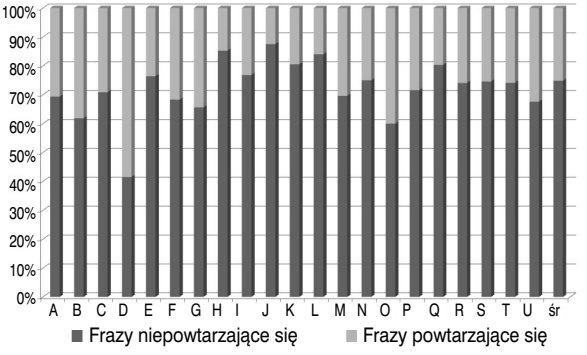
W przypadku „długiego ogona" słów kluczowych (fraz) dla wizyt z wyszukiwarek, najwięcej jest słów kluczowych, które osobno tworzą niewielki ruch na stronie. Razem zebrane, decydują o popularności witryny. „Długi ogon" dla tego zjawiska to ruch w witrynach generowany przez rzadko szukane słowa kluczowe, tzw. słowa z długiego ogona. W wielu wypadkach skuteczną metodą popularyzacji witryny w internecie jest koncentrowanie się na mniej konkurencyjnych słowach kluczowych, które zebrane razem tworzą wysoki udział w generowaniu ruchu w serwisie internetowym.

Budowa fraz kluczowych

Ostatnim mierzalnym parametrem, charakteryzującym wysoką widoczność w wynikach wyszukiwania, są formułowane zapytania do wyszukiwarki. Można zaproponować hipotezę, że ilu użytkowników wyszukiwarek, tyle sposobów na wyszukanie informacji. Autorzy Google podają[8], że około 50% wpisywanych do niej zapytań każdego dnia jest niepowtarzalnych, nigdy wcześniej niestworzonych.

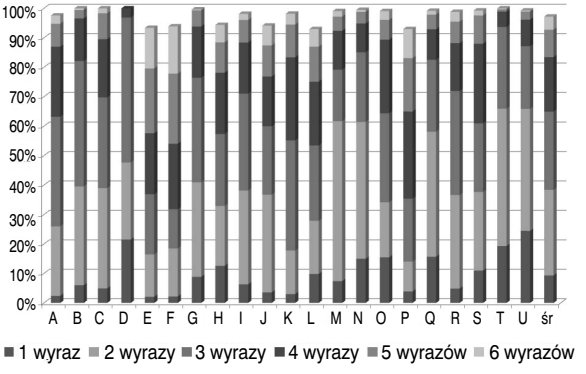
Na rysunku 4 widać zróżnicowany poziom niepowtarzalnych fraz kluczowych. Są one przedstawione w stosunku do całkowitej liczby wszystkich fraz, bez powtórzeń. Obraz zmieni się odwrotnie, gdyby brać pod uwagę powtórzenia, czyli całkowitą liczbę wizyt pochodzących z wyszukiwarek. Średnia ważona przedstawia łączny udział fraz niepowtarzających się ze wszystkich witryn.

Rysunek 4. Całkowity udział fraz niepowtarzalnych do powtarzających się



Źródło: opracowanie własne

Rysunek 5. Rozkład długości fraz kluczowych wpisywanych do wyszukiwarki



Źródło: opracowanie własne

W statystykach odwiedzin najczęściej powtarzają się kombinacje, dwu-, trzy- i czterowyrazowe. Rysunek 5 ilustruje rozkład długości wpisywanych fraz do wyszukiwarki. Osiem witryn zanotowało kilkukrotne wejścia dla fraz o długościach od 7 do 13 wyrazów, natomiast dwie frazy (J i L) od 14 do 20 słów kluczowych, ale nie zostały one pokazane na wykresie. Rysunek przedstawia liczbę różnych długości fraz wprowadzonych do wyszukiwarki, bez uwzględniania częstotliwości ich ponownego wprowadzenia. Opiera się on wyłącznie na jednokrotnych wystąpieniach. Średnia ważona przedstawia całkowity rozkład pochodzący ze wszystkich witryn.

Wnioski

Na podstawie przeprowadzonych studiów i analiz wskazano korzyści, jakie dają witrynom internetowym umiejętna budowa stron i wykorzystanie technologii

[7] http://www.zus.pl/default.asp?p=1&id=35, [21.07.2007].
[8] J. Battelle, Szukaj. *Jak Google i konkurencja wywołali biznesową i kulturową rewolucję*, Wydawnictwo PWN, Warszawa 2006, s. 27.

tekstowej. Powodują one znaczne zwiększenie widoczności witryny w wynikach wyszukiwania. Wyraźnie podkreślono, jak możliwa jest dywersyfikacja wizyt i ruchu pochodzącego z wyszukiwarek. Ruch nie powinien być oparty wyłącznie na kilku popularnych frazach ze słów kluczowych, a powinien uwzględniać frazy z „długiego ogona wyszukiwania". Widać to doskonale w witrynach składających się z wielu podstron. Nawet jeśli okresowo witryna pod pewnymi frazami ze słów kluczowych nie jest na wysokich pozycjach, to pozostają jeszcze możliwości pod innymi frazami. Słabiej na tym tle wypadają witryny z zawartością multimedialną lub małą liczbą podstron. Mogą być atrakcyjniejsze w wyglądzie, ale nie ma to absolutnie znaczenia, jeśli z powodu braku widoczności w wyszukiwarkach, w ogóle nie będą odwiedzane.

Bibliografia

Ch. Anderson, *The Long Tail: Why the Future of Business is Selling Less of More,* Hyperion, Nowy Jork 2006.

J. Battelle, *Szukaj. Jak Google i konkurencja wywołali biznesową i kulturową rewolucję*, Wydawnictwo PWN, Warszawa 2006.

A. Bifet, C. Castillo, A. Chirita, I. Weber, *An analysis of factors used in search engine ranking*, 1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), Chiba 2005.

M.P. Evans, *Analysing Google rankings through search engine optimization data*, „Internet Research" 2007, tom 17, wyd. 1.

S. Fortunato, M. Boguna, A. Flammini, F. Menczer, *How to make the top ten: approximating PageRank from In-degree,* Materiały z 14th International World Wide Web Conference, Edynburg 2006.

B.J. Jansen, A. Spink, *How are we searching the world wide web? A comparison of nine search engine transaction logs,* „Information Processing and Management" 2006, nr 42.

A. Khaki-Sedigh, M. Roudaki, *Identification of the dynamics of the Google ranking algorithm,* 13th IFAC Symposium On System Identification, Iran 2003.

L. Page, S. Brin, R. Motwani, I T. Winograd, *The PageRank Citation Ranking: Bringing Order to the Web,* Technical Report, Stanford University 1998.

Autor pracuje jako asystent w Katedrze Informatyki Akademii Ekonomicznej w Katowicach. Od niedawna zajmuje się problematyką promocji w internecie. Jego zainteresowania dotyczą tematyki marketingu w wyszukiwarkach internetowych oraz reklamy w internecie.

POLECAMY

***E-COMMERCE 2008. Najnowsze tendencje rynku handlu elektronicznego,* 16–17 października 2007 r., Warszawa**

W dniach 16 i 17 października odbędzie się w Warszawie konferencja *E-COMMERCE 2008 – Najnowsze tendencje rynku handlu elektronicznego*. Konferencja będzie poruszała tematy: efektywnej reklamy, handlu wirtualnego, społeczności wirtualnych, kwestii prawnych prowadzenia działalności e-handlowej, tendencji, które pojawią się na rynku handlu elektronicznego w ciągu najbliższych kilku lat.
Ponadto omówione zostaną najnowsze trendy sprzedażowe, wpływ personalizacji w internecie na podejmowanie decyzji przez klienta, a także sposoby budowania zaufania i lojalności klienta online.
Więcej informacji na: **http://www.informedia-poland.com/client/Index.aspx?id=conference&sub=introduction&conflD=463**

***IT Governance, Nowa Strategia Wsparcia Biznesu,* 18-19 października 2007 r., Warszawa**

IT Governance to koncepcja ładu korporacyjnego w obszarze IT (technologii informacyjnych), czyli usystematyzowanego podejścia do efektywnego zastosowania technologii informacyjnych w przedsiębiorstwie, którego celem jest maksymalizowanie wartości przedsiębiorstwa. Do głównych elementów ładu korporacyjnego w obszarze IT należą: jasne reguły podejmowania decyzji, jasne reguły inwestowania w IT, czytelna struktura zarządcza z zakresem odpowiedzialności. Koncepcja ta zyskuje znaczenie wraz ze wzrostem wielkości inwestycji w IT, a wprowadzenie odpowiednich zasad zarządzania, daje kadrze zarządzającej większą pewność, że podjęte inwestycje w IT w rzeczywisty sposób przyczyniają się do kreowania wartości biznesowej.
Informedia Polska organizuje w dniach 18 i 19 października kongres poświęcony zagadnieniom *IT Governance*. Wśród tematów przewidzianych przez organizatorów znajdują się:
– *Strategiczna rola IT w tworzeniu wartości firmy,*
– *Rola i znaczenie komunikacji w prawidłowej współpracy poziomu IT i poziomu biznesowego,*
– *Realizacja strategii korporacyjnej a zarządzanie IT – wzajemne zależności,*
– *Analiza długofalowych potrzeb biznesowych: klucz do integracji IT ze strategią firmy,*
– *IT Governance – wartość dodana czy kluczowy czynnik zwiększający wartość firmy?*
Więcej informacji na: **http://www.informedia-polska.pl/client/index.aspx?sub=introduction&id=conference&ConflD=419**