B A D A N I A O P E R A C Y J N E I D E C Y Z J E

Nr 3 2009

Ireneusz KACZMAR*

ZASTOSOWANIE MACIERZY PRZEPŁYWÓW MIĘDZYGAŁĘZIOWYCH DO WYZNACZANIA HIERARCHII WAŻNOŚCI INFROMACJI W INTERNECIE

Autor rozpatruje graf Internetu jako globalną gospodarkę, w której jedynym wytwarzanym produktem jest informacja. Założono, że w przestrzeni sieci występuje hierarchia ważności informacji i poszczególnych słów kluczowych. Ważne słowa kluczowe tworzą duże grafy informacji, które przyciągają słabsze grafy, składające się z mniej znaczących danych. Grafy informacji są wzajemnie powiązane jak gałęzie produkcji w klasycznej gospodarce. W związku z tym, że w sieci informacji jest nieskończenie dużo, trudno ją mierzyć jakościowo lub badać, czy jest fałszywa, czy prawdziwa. Zaproponowano nietypowy, ilościowy pomiar informacji oraz analizę wag, jakie zachodzą między poszczególnymi słowami kluczowymi. Wykorzystano wyniki z wyszukiwarek internetowych. Do analizy ilości informacji oraz zależności, jakie zachodzą między poszczególnymi słowami kluczowymi użyto macierzy przepływów międzygałęziowych.

Słowa kluczowe: *graf Internetu, produkcja informacji, wektor, graf informacyjny, model przepływów międzygałęziowych, analiza słów kluczowych*

Wprowadzenie

Z ekonomicznego punktu widzenia informacja jest towarem. Na pewno więc ważniejsza informacja (tak jak lepsza książka) jest droższa od tej mniej ważnej. Wszystko zależy od tego, co kogo interesuje, a większy popyt jest zwykle na to, co interesuje większość. Dziś nieograniczonym źródłem informacji jest Internet. Jakich informacji w sieci jest zatem więcej, czym się interesujemy i czego poszukuje większość? Jak zmierzyć ilość informacji i zbadać relacje zachodzące pomiędzy słowami kluczowymi,

* Dział Rozwoju Uczelni, Państwowa Wyższa Szkoła Wschodnioeuropejska w Przemyślu, ul. Tymona Terleckiego 6, 37-700 Przemyśl, e-mail: dru@pwsw.pl

wreszcie jak ważne są poszczególne słowa w sieci dla jej użytkowników? Oto podstawowe pytania, na które – może przynajmniej częściowo – odpowie niniejsze opracowanie.

Jeżeli występuje jakaś hierarchia ważności informacji w Internecie, to poszczególne słowa kluczowe również mogą mieć swoją cenę. Tak jest istotnie. Dla reklamodawców, zlecających np. umieszczanie linków sponsorowanych, wyszukiwarki drożej wyceniają słowa kluczowe z branży np. biznes i praca w porównaniu z branżą hobby czy rozrywka. Trudniej też pozycjonuje się w wyszukiwarkach witryny zawierające popularne słowa, ponieważ jest ich po prostu więcej (tworzą większy graf informacyjny). W erze rewolucji informacyjnej integralną częścią życia człowieka jest wiele rozmaitych komunikatów, które ciągle do nas docierają. Zanurzeni w nurcie wiadomości, nie zawsze wiemy, jak dobrze wykorzystywać zasoby informacyjne Internetu. Coraz większy natłok różnych danych powoduje często brak ich logicznego umiejscowienia, czy też właściwego powiązania. Wówczas niezbędny staje się system, który pomoże podjąć właściwą decyzję. Koncentrując się na najnowszych technologiach elektronicznego przepływu danych, kładzie się dziś nacisk na globalną społeczność informacyjną, pozostającą w zasięgu rozległej, nieograniczonej sieci teleinformatycznej. Integracja nowoczesnych rozwiązań w łączności, elektronice i informatyce uwidacznia nam codziennie, że jesteśmy cząstką jednego wielkiego systemu, w którym komunikacja między poszczególnymi jednostkami – ludźmi, niezależnie od odległości, nie nastręcza już żadnych trudności.

Nie bez znaczenia jest również niematerialna postać oraz fizyczna forma reprezentacji informacji. Jak ją zmierzyć, skoro nie można jej dotknąć ani zobaczyć. Metody i środki użyte do pomiaru i analizy relacji zachodzących w zbiorach informacyjnych muszą być więc odpowiednie. W pierwszej części artykułu przybliżona zostanie czytelnikowi definicja, rodzaje i formy reprezentacji informacji. Następnie przedstawiono klasyczny model przepływów międzygałęziowych, stosowany do analizy relacji zachodzących w tradycyjnej gospodarce. W ostatniej części artykułu opisano sposób budowania grafu i wektora informacji oraz zastosowanie macierzy przepływów międzygałęziowych do ustalania hierarchii ważności słów kluczowych w sieci.

1. Definicja informacji i jej rodzaje

Informacja jest niematerialnym dobrem, którego znaczenia bardzo często nie doceniamy. Otacza nas, ale nie zwracamy uwagi jak, kiedy, i skąd do nas dociera. Dla człowieka jest czymś naturalnym, ponieważ w naszym życiu następuje ciągły dopływ nowych danych oraz ich wymiana. Samo pojęcie informacji nie jest w pełni definiowalne, podobnie jak w matematyce aksjomat (np. punkt), a rozważania o jej istocie możemy przeprowadzać tylko w określonym kontekście. Krótko mówiąc, zinterpretowana lub

przetworzona dana, której nadano określone znaczenie jest już informacją, np. 60 – dana liczbowa, ale 60 km/h – prędkość, informuje już o czymś konkretnie. Komunikat jest zestawem informacji (liczbowych, tekstowych, graficznych lub w innej formie), stanowiących przedmiot przekazu (wymiany) między dowolnymi komunikującymi się partnerami. Informacja lub komunikat mogą mieć dowolny charakter, między innymi gospodarczy, kiedy nadamy jej znaczenie gospodarcze lub ekonomiczne. Może być przedmiotem aktu kupna–sprzedaży, stanowić towar. Występuje także jako czynnik produkcji, wiedza + ziemia, praca, kapitał = siła wytwórcza w społeczeństwie informacyjnym. Stanowi zasób gospodarki niematerialnej, co nie znaczy jednak, że nie może mieć wymiernej wartości. Sposób jej przekazu, odbioru czy kodowania jest już sprawą umowną, musi jednak istnieć język lub inny sposób, pozwalający na komunikację między jednostkami zainteresowanymi wymianą danych.

Możemy również przyjąć, że informacja jest to mniej lub bardziej szczegółowe sprawozdanie (relacja) z jakiegoś faktu, należące do jak najbardziej bezpośredniej teraźniejszości. Istotne jest, aby sprawozdanie było bezstronne. Nie ma tutaj znaczenia żadna definicja, ważne jest praktyczne posługiwanie się informacją, umiejętność jej wydobycia, przekazania czy określenia warunków, jakie musi spełniać, aby była wartościowa. Aby informacja była naprawdę wyczerpująca, musi spełniać określone warunki, powinna odpowiadać przede wszystkim na pytania: co? gdzie? kiedy? jak? dlaczego? Jeśli brakuje jakiegoś elementu, informacja jest niepełna, sprzeczność tych elementów może spowodować **chaos** informacyjny, z którego trudno cokolwiek wyłowić. Kolejność odpowiedzi na wyżej postawione pytania uwarunkowana jest tym, co w danym wydarzeniu, sprawozdaniu czy komunikacie jest dla nas najważniejsze.

Informacja jest elementem wiedzy, faktem, wiadomością, komunikatem lub wskazówką, gromadzoną, komunikowaną lub przekazywaną komuś za pomocą jakiegoś kodu lub języka [6]. Podstawowe cechy informacji możemy wyszczególnić w punktach:

- stanowi pojęcie pierwotne, definiowalne tylko w określonym kontekście;
- ma charakter niematerialny i różne formy;
- zmniejsza stopień niewiedzy o badanym zjawisku;
- polepsza znajomość otoczenia, zaspokajając nasze potrzeby informacyjne;
- może być zasobem produkcyjnym, przejawiać charakter ekonomiczny, gospodarczy itp.;
- jest elementem wiedzy umożliwiającej budowanie systemów informacyjnych;
- stanowi wartość subiektywną (indywidualna waga, indywidualna interpretacja przez każdego człowieka czy organizację);
- może mieć różne źródła pochodzenia (element komunikatu, sprawozdania);
- może być prawdziwa lub fałszywa.

Jak mówił Wiener, informacja nie jest ani materią, ani energią, jest ona bowiem w naszym rozumieniu bardziej powiązana ze świadomością, będącą atrybutem istot myślących. Powinna więc znajdować się gdzieś w obszarze pomiędzy fizyką a psycho-

logią jako odzwierciedlenie otaczającej nas fizycznej rzeczywistości. Procesy informacyjne natomiast można zaliczyć do procesów wiążących człowieka z otoczeniem. Wiadomość przekazujemy w celu zmiany świadomości u jakiegoś osobnika, wydatkujemy przy tym zawsze określoną ilość energii – w zależności od tego, jaki sposób przekazu wybieramy. Najciekawsze jest to, iż informacja nie jest energią ani materią, a może przynosić materialne korzyści i bez energii nie może istnieć ani być przekazywana [7].

Również analogie termodynamiczne były często wykorzystywane w klasycznej teorii informacji. Pojawia się tutaj pojęcie entropii, czyli miary nieokreśloności. Według teorii Boltzmanna entropia gazu zmienia się w tym kierunku co liczba stanów, a więc osiąga maksimum, gdy informacja jest minimalna. Można więc powiedzieć, że:

- zerowa entropia to pełna informacja,
- wielka entropia to informacja zerowa.

Jak wiadomo z kinetycznej teorii gazów, nawet w idealnym ciele, które jest gazem w izolowanym naczyniu, panuje molekularny chaos. Nie jesteśmy w stanie prześledzić wszystkich ruchów cząsteczek. Możemy jedynie określić podstawowe wielkości fizyczne, takie jak temperatura, ciśnienie i inne, czyli makroskopowy stan gazu. Stanu mikroskopowego na poziomie drgań elementarnych cząsteczek nie możemy określić, ponieważ w strukturze ciała panuje nieustanny ruch. Im większa jest ilość stanów cząsteczek, tym mniejsza informacja i wielka entropia. Możemy określić jedynie stopień naszej niewiedzy poprzez obliczenie liczby możliwych stanów mikroskopowych realizujących dany stan makroskopowy, nazwany prawdopodobieństwem termodynamicznym. Nie odpowiedziano jednak ostatecznie na pytanie, czy porównania termodynamiczne pomagają w teorii informacji, ponieważ wielu uważa, że problem informacji jest ogólnonaukowy i nie należy go rozwijać w jakiejś wąskiej dziedzinie przedmiotowej.

Potocznie używa się wyrazu „informacja” w znaczeniu wiadomość, komunikat itp., określając w ten sposób zarówno produkt działania informacyjnego, jak i samo działanie. W początkach XX wieku wzrosła rola informacji i zainteresowanie społeczne treścią wyrazu „informacja” oraz możliwością jej mierzenia. Treść próbowano powiązać z pojęciem „prawdopodobieństwo” [P. Fischer, 1921], lub z pojęciem „entropia” [L. Szillard, 1929]. W 1928 roku Hartley zaproponował logarytmiczną miarę informacji, a w 1948 roku Shannon uzasadnił matematyczny opis informacji jako miary zmniejszania nieokreśloności (niepewności), nadając tym samym pojęciu „informacja” być może pewien sens heurystyczny. Późniejsze dyskusje wykazały, że informacja jest częścią pojęcia „prawdopodobieństwo”, a odwrotnie – prawdopodobieństwo jest częścią informacji, podobnie jak entropia, za pomocą której można opisać pewne procesy informacyjne [7]. Dyskusja o naturze informacji nadal trwa i na pewno powinna być kontynuowana, gdyż tylko w ten sposób można będzie znaleźć najlepsze określenie tego pojęcia. W teorii informacji możemy wyróżnić dwa zasadnicze podejścia:

• Ilościowe [Shannon, 1948] – ilość informacji I zawarta w komunikacie B o zdarzeniu U równa jest różnicy pomiędzy początkową niepewnością zdarzenia U a niepewnością, jaka pozostaje na temat zdarzenia U po nadejściu komunikatu B.

• Jakościowe [Langefors, 1973] – uwypuklony jest tu aspekt semantyczny (znaczeniowy) danych, występują symbole, takie jak: dane, wiadomości, informacje i rekordy; danymi są zestawy sygnałów emitowanych przez otoczenie celowo lub nie i przyjmowanych przez odbiorcę. Wiadomość określa treść danych, jakie odbiorca jest w stanie wydobyć. Informacja – wiadomość, która zmniejsza niewiedzę odbiorcy. Rekord jest pojedynczym zestawem danych reprezentujących komunikat (wiadomość).

Wraz ze słowem „informacja" bardzo często pojawiają się terminy bliskoznaczne, takie jak:

`Sygnał` – jakaś zmienna w czasie, zjawiska fizyczne występujące w określonym przedziale czasu i zlokalizowane w konkretnym punkcie przestrzeni.

`Znak` – sygnał elementarny, przyporządkowany pewnej elementarnej treści.

`Wiadomość` – ciąg znaków.

`Komunikat` – wiadomość z logicznie uporządkowaną treścią w jakimś określonym języku.

`Informacja` – szeroko rozumiana wiadomość.

`Język` – zbiór symboli i reguł służących do komunikacji.

Analogie lingwistyczne

ZNAK	–	SYMBOL
ZBIÓR ZNAKÓW	–	ALFABET
SZEREG ZNAKÓW	–	SŁOWO, WYRAZ, ZDANIE
LOGIKA TWORZENIA CIĄGÓW	–	GRAMATYKA
KOMUNIKAT	–	ZDANIE SENSOWNE
ZBIÓR ZNAKÓW + LOGIKA	–	JĘZYK

2. Klasyczny model Leontiewa

Znajomość modelu Leontiewa jest niezbędna do zrozumienia dalszej części artykułu. Jest on znany także pod nazwami: model przepływów międzygałęziowych, model „input-output" czy model nakładów i wyników. Jego twórcą jest amerykański uczony Wassily Leontiew, laureat nagrody Nobla w dziedzinie ekonomii w 1973 roku za *For the*

development of the input-output method and for its application to important economic problems. Model ten daje możliwość opisywania i analizy złożonych systemów gospodarczych. Opiera się na obserwacji, że w skład gospodarki wchodzi wiele gałęzi produkcyjnych, których działalność jest wzajemnie powiązana. Powiązania te wynikają stąd, że produkcja jednych gałęzi jest zużywana jako nakład w innych gałęziach. Dodatkowo część produkcji zostaje przeznaczona na zaspokojenie potrzeb odbiorców końcowych (sektora gospodarstw domowych czy tworzenia zapasów). Model Leontiewa umożliwia odpowiedź na pytanie: jaka powinna być produkcja każdej gałęzi gospodarki, aby zrównoważyć popyt zgłaszany zarówno przez same gałęzie, jak i sektor gospodarstw domowych. Pozwala również na analizę zmian w strukturze produkcji, które są wywołane zmianami zapotrzebowania ze strony sektora gospodarstw domowych lub wielkości produkcji jednej z gałęzi [5]. Zwykle analiza obejmuje wiele gałęzi i ma dość skomplikowaną strukturę, aby więc przenieść mechanizm na grunt zależności, jakie zachodzą w Internecie, przyjmujemy pewne założenia:

- całkowity (globalny) poziom produkcji ilości informacji w Internecie dla każdego podgrafu informacyjnego jest uzależniony od wzajemnych powiązań między słowami kluczowymi;
- każdy podgraf informacyjny wytwarza jeden typ słowa kluczowego lub grupę słów kluczowych w stałych proporcjach i do tego potrzebuje jedno słowo lub grupę słów kluczowych, również w stałych proporcjach, z innego podgrafu informacyjnego;
- sektor gospodarstw domowych tworzący wektor [**d**], który obrazuje ważność danego słowa kluczowego względem innych w rozpatrywanej sekwencji słów, pokazuje popyt na informację wśród użytkowników przestrzeni Internetu.

Załóżmy, że gospodarka składa się z n gałęzi produkcyjnych (mogą to być również sektory czy działy pojedynczej firmy). Wprowadzamy następujące oznaczenia, wyrażone w jednostkach pieniężnych:

- X_i (i = 1, 2, ..., n) – wielkość produkcji całkowitej (globalnej) i-tej gałęzi,
- x_{ij} (i, j = 1, 2, ..., n) – część produkcji i-tej gałęzi, która jest zużywana (przepływa) na potrzeby produkcji gałęzi j-tej,
- d_i (i = 1, 2, ..., n) – produkt końcowy i-tej gałęzi (różnica między produkcją całkowitą i-tej gałęzi a jej przepływami do wszystkich gałęzi).

Tabela 1. Tablica przepływów międzygałęziowych

Numer gałęzi		Przepływy x_{ij}				Produkt końcowy	Produkcja całkowita
		j					
		1	2	...	n	d_i	X_i
i	1	x_{11}	x_{12}	...	x_{1n}	d_1	X_1
	2	x_{21}	x_{22}	...	X_{2n}	d_2	X_2
	⋮	⋮	⋮	⋱	⋮	⋮	⋮
	n	x_{n1}	x_{n2}	...	x_{nn}	d_n	X_n

Punktem wyjścia modelu Leontiewa jest bilans gospodarczy w postaci tablicy przepływów międzygałęziowych (zob. tab. 1), przygotowany w sposób umożliwiający kwantyfikacje wzajemnych powiązań między wyodrębnionymi częściami systemu. Tablica zawiera dane liczbowe, charakteryzujące działalność gospodarczą w pewnym okresie czasowym.

Ponieważ produkcja całkowita gałęzi *i*-tej jest sumą przepływów międzygałęziowych oraz produktu końcowego, otrzymujemy układ równań bilansowych postaci:

$$\begin{cases} \mathbf{X}_1 = x_{11} + x_{12} + ... + x_{1n} + \mathbf{d}_1 \\ \mathbf{X}_2 = x_{21} + x_{22} + ... + x_{2n} + \mathbf{d}_2 \\ \quad ..... \\ \mathbf{X}_n = x_{n1} + x_{n2} + ... + x_{nn} + \mathbf{d}_n \end{cases}. \tag{1}$$

Na mocy założenia o stałych proporcjach zużywanej produkcji gałęzi *i*-tej przez gałąź *j*-tą możemy określić współczynniki

$$a_{ij} = \frac{x_{ij}}{\boldsymbol{X}_j}, \quad (i, j = 1, 2, ..., n), \tag{2}$$

nazywane współczynnikami kosztów. Macierz $\mathbf{A} = [a_{ij}]$ nazywamy macierzą współczynników kosztów. Współczynniki a_{ij} przyjmują wartości ze zbioru [0, 1] i są interpretowane następująco: aby w *j*-tej gałęzi uzyskać produkcję całkowitą o wartości jednej jednostki pieniężnej, należy zużyć produkcje gałęzi *i*-tej o wartości a_{ij} jednostek pieniężnych. Z zależności (2) otrzymujemy

$$x_{ij} = a_{ij}\mathbf{X}_j, \quad (i, j = 1, 2, ..., n), \tag{3}$$

co pozwala zapisać układ (1) w postaci:

$$\begin{cases} \mathbf{X}_1 = a_{11}\mathbf{X}_1 + a_{12}\mathbf{X}_2 + ... + a_{1n}\mathbf{X}_n + \mathbf{d}_1 \\ \mathbf{X}_2 = a_{21}\mathbf{X}_1 + a_{22}\mathbf{X}_2 + ... + a_{2n}\mathbf{X}_n + \mathbf{d}_2 \\ ..... \\ \mathbf{X}_n = a_{n1}\mathbf{X}_1 + a_{n2}\mathbf{X}_2 + ... + a_{nn}\mathbf{X}_n + \mathbf{d}_n \end{cases}. \tag{4}$$

To z kolei umożliwia zapisanie układu równań bilansowych (4) w postaci macierzowej:

$$\begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_n \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & ... & a_{nn} \end{pmatrix} \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_n \end{pmatrix} + \begin{pmatrix} \mathbf{d}_1 \\ \mathbf{d}_2 \\ \vdots \\ \mathbf{d}_n \end{pmatrix} \tag{5}$$

lub w postaci skróconej

$$\mathbf{X} = \mathbf{AX} + \mathbf{d}, \tag{6}$$

gdzie **X** oznacza macierz (wektor) produkcji całkowitej (globalnej), **A** macierz współczynników kosztów, **d** – macierz (wektor) produktu końcowego. Równanie (6) zapisujemy w postaci tzw. modelu Leontiewa, gdzie **I** oznacza macierz jednostkową stopnia *n*.

$$\mathbf{X} - \mathbf{AX} = \mathbf{d} \Leftrightarrow (\mathbf{I} - \mathbf{A})\mathbf{X} = \mathbf{d}\,. \tag{7}$$

Macierz (**I** − **A**) nosi nazwę macierzy Leontiewa i przekształca wektor produkcji całkowitej **X** w wektor produktu końcowego **d**. Powstaje natychmiast pytanie, czy znając wektor produktu końcowego **d**, możemy odwrócić sytuację i wyznaczyć wektor produkcji całkowitej **X**? Aby na nie odpowiedzieć, wprowadzamy pojęcie macierzy produktywnej. Macierz **A** współczynników kosztów jest produktywna, jeżeli istnieje nieujemny wektor produkcji całkowitej **X**, taki ze **X** > **AX**. Z ekonomicznego punktu widzenia oznacza to, że musi istnieć chociaż jeden wektor produkcji całkowitej, przy którym produkcja całkowita przewyższa zużycie produkcyjne (przepływy międzygałęziowe). Gdyby taki wektor nie istniał, oznaczałoby to, że gospodarka nie jest w stanie wytworzyć w każdej gałęzi więcej niż zużywa na potrzeby bieżącej produkcji, czyli byłaby to gospodarka „zjadająca sama siebie". Z tego względu w realnej gospodarce możemy założyć, że macierz **A** jest produktywna. Zastosowanie mają dwa następujące twierdzenia.

Twierdzenie 1. *Jeżeli macierz* **A** *jest produktywna*, *to macierz Leontiewa* (**I** − **A**) *jest macierzą nieosobliwą*.

Twierdzenie 2. *Jeżeli macierz* **A** *jest produktywna*, *to wszystkie elementy macierzy* $(\mathbf{I} - \mathbf{A})^{-1}$ *są nieujemne*.

Z twierdzenia 1 wynika bezpośrednio, że w realnej gospodarce produkt końcowy **d** wyznacza w sposób jednoznaczny produkcję całkowitą **X** zgodnie z regułą

$$\mathbf{X} = (\mathbf{I} - \mathbf{A})^{-1}\mathbf{d}\,. \tag{8}$$

Dodatkowo z twierdzenia 2 wynika, że dla dowolnego nieujemnego wektora produktu końcowego **d** otrzymamy również nieujemny wektor produkcji całkowitej **X** [5].

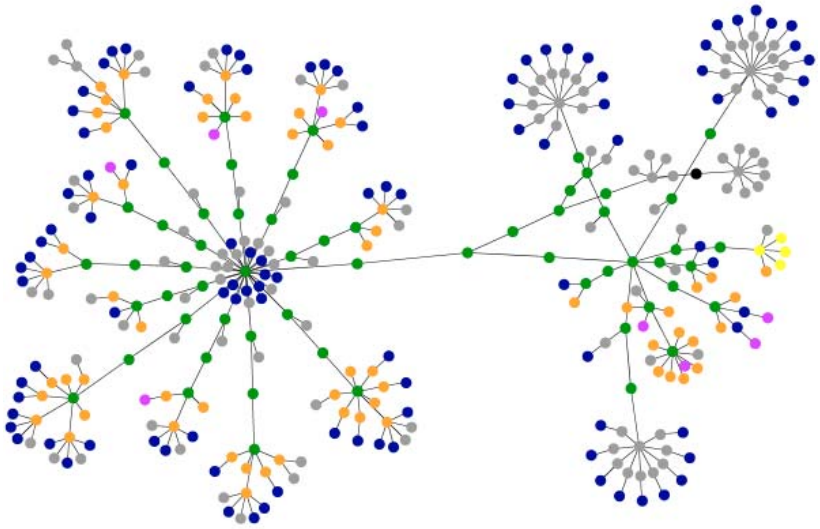
Znajomość macierzy przepływów międzygałęziowych pozwoli zrozumieć filozofię wykorzystania tego modelu do analizy informacji w grafie Internetu, którą zaproponowano w dalszej części artykułu. Globalna sieć zostanie porównana do globalnej gospodarki, gdzie gałęzie przemysłu będą reprezentowane przez grafy informacji, skupione wokół określonych słów kluczowych. Ważniejsze słowa kluczowe będą tworzyły większe grafy informacji, tak jak ważne dla gospodarki gałęzie przemysłu.

Mniej ważne słowa kluczowe utworzą mniejsze podgrafy informacji, ponieważ jest ich mniej w sieci i w hierarchii ważności znajdują się niżej. W końcowej części podano również przykłady liczbowe.

3. Grafy informacji w Internecie

Trudno rozpatrywać sieć globalną całościowo. Witryny internetowe tworzą jeden wielki graf informacyjny, zawierający nieskończoną ilość słów kluczowych we wszystkich językach świata. W dotychczas spotykanych opracowaniach autorzy określali pojedyncze strony www jako węzły sieci, natomiast połączenia *url* jako jego wierzchołki. W uproszczeniu witrynę (cały serwis internetowy), liczącą kilkadziesiąt pojedynczych stron, można zobrazować rysunkiem 1, a miliony takich witryn tworzy wielki graf Internetu.

Oczywiście graf Internetu jest o wiele bardziej rozbudowany i skomplikowany niż przedstawiony na rysunku. Obecnie liczy ponad kilka bilionów stron www, zaindeksowanych w wyszukiwarkach. Pozostają jeszcze strony niezaindeksowane i te, które nie są wyświetlane; dokładnej informacji co do wielkości grafu Internetu nikt nie podaje. W tej tradycyjnej reprezentacji graficznej serwisu internetowego kropki (węzły) oznaczają pojedyncze strony www, a połączenia *url* są liniami łączącymi.



Rys. 1. Graf witryny internetowej
Ź r ó d ł o: Opracowania własne.

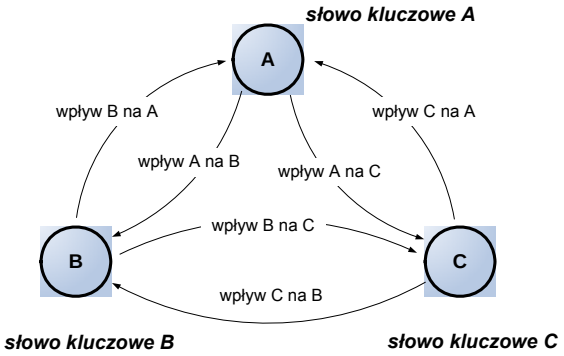
Jak już wspomniano, sieć Internetu jest zbyt duża, aby analizować relacje zachodzące między wszystkimi jego węzłami. Nawet jeśli przyjmiemy, że pojedyncza strona www jest węzłem, to nie ma takiego komputera, który by sprawdzał wszystkie węzły w sensownym czasie. Stron internetowych i tak będzie przybywać w szybszym tempie niż będzie zwiększała się wydajność sprzętu. Pojedyncza strona www to tylko plik z rozszerzeniem .html czy .php, który nie ma takiego znaczenia dla przestrzeni jak treść w nim zawarta. W interakcję z przestrzenią wchodzi treść, czyli słowa kluczowe zawarte przede wszystkim w nagłówku i opisie, bo na to głównie reagują wyszukiwarki. Dlatego lepiej analizować relacje zachodzące w podgrafach informacyjnych w obrębie wybranych słów kluczowych (węzłów). W niniejszej pracy zastosowano takie właśnie podejście. Założono, że węzłem grafu informacyjnego jest nie pojedyncza strona internetowa, ale pojedyncze słowo kluczowe np. *słowo x*, *słowo y*, *słowo z*. Rozpatrywany graf będzie więc miał tyle węzłów, ile słów kluczowych będziemy brali pod uwagę. Możemy budować dowolnie duży graf informacyjny, składający się z takiej ilości słów kluczowych, na jaką pozwoli moc obliczeniowa komputera.

Traktowanie pojedynczej strony www jako węzła sieci nie jest dobre w tym przypadku. Byłoby prawidłowe, gdyby strona www składała się tylko z jednego słowa kluczowego, a przecież strony www zawierają w nagłówku i opisie co najmniej po kilka słów. W związku z tym mogą być znajdywane przez wyszukiwarki po różnych zapytaniach. Wchodzą więc w interakcję z przestrzenią w zależności od kontekstu zapytania, po jakim szukamy danej informacji. Na przykład strona internetowa wyższej uczelni może być znaleziona w sieci poprzez zapytania o słowa: *szkoła wyższa*, *uczelnia*, *politechnika*, *uniwersytet*, *wyższa zawodowa* ... itd. Jest to więc sześć słów kluczowych, będących węzłami grafu, uruchamiającymi się w zależności od tego, czy o nie pytamy, czy nie. Jak się okazuje, mamy więc sześć węzłów na jednej stronie internetowej, a nie jedną stronę internetową jako jeden węzeł grafu, w którym jest sześć słów.

Jeżeli węzłem grafu informacyjnego jest słowo kluczowe, to wierzchołkiem grafu niech będzie odpowiedź wyszukiwarki internetowej na to słowo. Wagą wierzchołka jest ilość odpowiedzi. Jeżeli zapytamy wyszukiwarkę o *słowo kluczowe x*, to otrzymamy na przykład 1 milion odpowiedzi, co oznacza, że nasza wyszukiwarka znalazła wierzchołek z wagą 1.000.000 dla węzła lub węzłów tym słowem związanych. Przyjmujemy wówczas wagę dla danego wierzchołka = 1.

Uwaga: dla wygody obliczeń i ze względu na to, że obecnie ilość odpowiedzi z wyszukiwarek jest bardzo duża, obcinamy zawsze sześć ostatnich zer. Ta zasada będzie obowiązywać w dalszej części opracowania.

Aby lepiej zrozumieć pojęcie grafu i wektora informacji, wystarczy zapoznać się z rysunkiem 2. Posłużymy się tutaj przykładem. Zapytamy najpopularniejszą wyszukiwarkę świata (w Polsce – google.pl) o trzy różne słowa kluczowe: x = *praca*, y = *nauka*, z = *polityka*.



Ilość odpowiedzi [w mln. stron www]
udzielonych przez wyszukiwarkę
na zapytanie o słowo kluczowe x, y, z.

$$\text{Wektor ilości informacji} = \begin{bmatrix} 79{,}5 \\ 42{,}1 \\ 31{,}2 \end{bmatrix}$$

Rys. 2. Graf i wektor informacji dla trzech słów kluczowych x, y, z.
Źródło: Opracowania własne.

8 października 2008 roku w odpowiedzi na słowo x otrzymujemy ponad 79 milionów odpowiedzi, dla słowa y – 42,1 miliona odpowiedzi, dla słowa kluczowego z – 31,2 miliona odpowiedzi. Świadczy to o tym, że sieć najwięcej informacji generuje dla słowa kluczowego x – *praca.* Jest to więc duży i silny graf informacyjny, tak jak silna gałąź gospodarki, np. budownictwo. Jeżeli więc słowa kluczowe porównamy do gałęzi produkcyjnych klasycznej gospodarki, to:

- wektor ilości informacji generowany przez sieć Internet dla wymienionych wcześniej trzech słów kluczowych będzie wyglądał analogicznie jak w modelu Leontiewa wektor produkcji całkowitej **X**:

$$\text{Wektor ilości informacji} = \begin{bmatrix} \text{Ilość odpowiedzi wyszukiwarki na zapytanie } x \\ \text{Ilość odpowiedzi wyszukiwarki na zapytanie } y \\ \text{Ilość odpowiedzi wyszukiwarki na zapytanie } z \end{bmatrix}.$$

Teraz, mając graf i wektor informacji, możemy przeanalizować relacje zachodzące między słowami kluczowymi x, y, z. Wystarczy zbudować układ równań liniowych i – stosując np. algebrę macierzy – znaleźć rozwiązanie. Przydatne będą wzory znane z modelu przepływów międzygałęziowych. Pozostaje jeszcze kwestia interpretacji wyniku, która zostanie przedstawiona później.

4. Zastosowanie macierzy przepływów międzygałęziowych do wyznaczania ważności słów kluczowych w sieci

Jeżeli Internet potraktujemy jak globalną gospodarkę, w której jedynym wytwarzanym produktem jest informacja, to musi występować jakaś hierarchia ważności

i wzajemne oddziaływanie tejże. Jak sprawdzić, które słowa kluczowe są ważniejsze? Zakładamy, że w produkcji informacji słowa kluczowe w jakiś sposób na siebie oddziałują, a jedna informacja może generować powstawanie następnej. Intuicyjnie można stwierdzić, że ważniejsze są słowa kluczowe, których jest więcej, ponieważ więcej informacji wytwarza sieć wokół słów częściej poszukiwanych przez internautów. Trudniej przecież spowodować, aby strona internetowa ze słowem kluczowym *praca* znalazła się na pierwszych miejscach wyników wyszukiwania, niż strona internetowa ze słowem kluczowym np. *herbata* w tytule. Strona internetowa z hasłem *praca* musi pokonać aż 79,5 miliona konkurentów, natomiast serwis internetowy opisujący herbatę musi pokonać zaledwie 3,5 miliona innych witryn, aby znaleźć się na pierwszym miejscu w wyszukiwarce. Wniosek nasuwa się sam: w grafie informacji – *praca,* liczącym ponad 79 milionów stron, panuje znacznie większa konkurencja niż w grafie z hasłem *herbata*, który liczy 3,5 miliona stron. W związku z tym w Internecie słowo *praca* na pewno jest ważniejsze niż słowo *herbata.* Dalej przyjmujemy założenia podobne jak w modelu przepływów międzygałęziowych:

- całkowity (globalny) poziom produkcji ilości informacji w Internecie dla każdego podgrafu informacyjnego jest uzależniony od wzajemnych powiązań między słowami kluczowymi;
- każdy podgraf informacyjny wytwarza jeden typ słowa kluczowego lub grupę słów kluczowych w stałych proporcjach i do tego potrzebuje jedno słowo lub grupę słów kluczowych, również w stałych proporcjach, z innego podgrafu informacyjnego;
- sektor gospodarstw domowych tworzący wektor [**d**] i obrazuje ważność danego słowa kluczowego względem innych w rozpatrywanej sekwencji słów, pokazuje popyt na informację wśród użytkowników przestrzeni Internetu.

Rozpatrzymy teraz zależności występujące pomiędzy słowami kluczowymi, przytoczonymi w poprzednim rozdziale. Wyszukiwarka google.pl, w odpowiedzi na słowa kluczowe: x = *praca*, y = *nauka*, z = *polityka*, daje (produkuje) następujące ilości informacji:

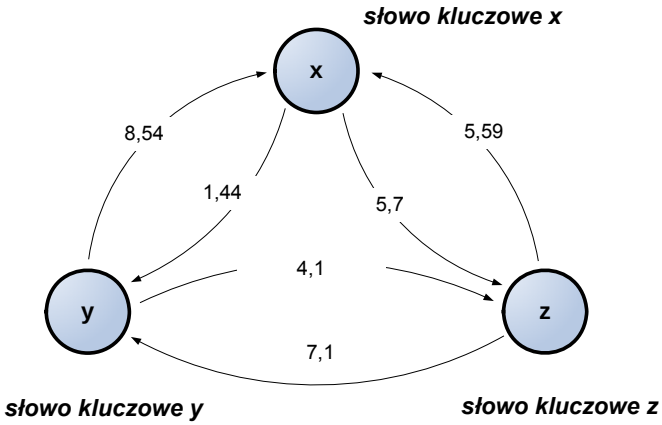
1. $x = 79{,}5$ miliona odpowiedzi,
2. $y = 42$ miliony odpowiedzi,
3. $z = 31{,}2$ milionów odpowiedzi.

Wektor ilości informacji przedstawia się więc następująco:

$$\mathbf{X} = \begin{bmatrix} 79{,}5 \\ 42 \\ 31{,}2 \end{bmatrix} \Rightarrow \begin{bmatrix} \text{praca} \\ \text{nauka} \\ \text{polityka} \end{bmatrix}.$$

Budujemy graf zależności informacyjnej dla interesujących nas słów kluczowych. Graf jest analogiczny jak w modelu przepływów międzygałęziowych. Węzły to słowa kluczowe x, y, z, natomiast wierzchołki to odpowiedzi na wybrane sekwencje. Jak już wcześniej wspomniano, zawsze obcinamy sześć ostatnich zer, aby uprościć obliczenia.

Ilość informacji w Internecie zmienia się dość dynamicznie i znacznie wyprzedza rzeczywistość, dlatego warto zaznaczyć, że dane zebrano 8 października 2008 roku z wyszukiwarki google.pl. Zależności zilustrowano na rysunku 3.



Rys. 3. Graf zależności informacyjnej.
Ź r ó d ł o: Opracowanie własne.

Na grafie z rysunku 3 możemy zaobserwować ciekawe zależności. W rozpatrywanym układzie zależności słowo *y* ma duży wpływ na wytwarzanie przez sieć globalną słowa *x*, a słowo kluczowe *z* znacznie oddziałuje na *y*, itd. Nie zawsze zależności występujące w sieci są zgodne z rzeczywistością, ale wszystko zależy od umiejętności właściwego doboru słów kluczowych, co zostanie pokazane w następnych przykładach. Na podstawie grafu budujemy macierz zależności informacyjnej, która jest analogiczna z macierzą sąsiedztwa, znaną powszechnie z teorii grafów.

$$\textbf{Macierz zależności informacyjnej} = \begin{bmatrix} \mathbf{0} & \mathbf{1,44} & \mathbf{5,7} \\ \mathbf{8,54} & \mathbf{0} & \mathbf{4,1} \\ \mathbf{5,59} & \mathbf{7,1} & \mathbf{0} \end{bmatrix}.$$

Na podstawie wzoru (2) z rozdziału drugiego macierz zależności informacyjnej przekształcamy w macierz współczynników informacji **A**. W modelu Leontiewa macierz $\mathbf{A} = [\mathbf{a}_{ij}]$ nazywano macierzą współczynników kosztów lub współczynników technicznych.

$$\textbf{Macierz współczynników informacji A} = \begin{bmatrix} \mathbf{0} & \mathbf{0,034} & \mathbf{0,182} \\ \mathbf{0,107} & \mathbf{0} & \mathbf{0,131} \\ \mathbf{0,07} & \mathbf{0,169} & \mathbf{0} \end{bmatrix}.$$

Następnie obliczamy macierz (**I** − **A**), w naszym przypadku przekształca ona wektor produkcji całkowitej informacji w sieci **X**, w wektor popytu końcowego **d**, informujący także o ważności słów kluczowych w tym wektorze. Macierz (**I** – **A**) oraz jej odwrotność wygląda następująco:

$$(\mathbf{I}-\mathbf{A})=\begin{bmatrix} \mathbf{1} & \mathbf{-0{,}0343} & \mathbf{-0{,}1827} \\ \mathbf{-0{,}1074} & \mathbf{1} & \mathbf{-0{,}1314} \\ \mathbf{-0{,}0703} & \mathbf{-0{,}169} & \mathbf{1} \end{bmatrix}, \quad (\mathbf{I}-\mathbf{A})^{-1}=\begin{bmatrix} \mathbf{1{,}0211} & \mathbf{0{,}0681} & \mathbf{0{,}1955} \\ \mathbf{0{,}1218} & \mathbf{1{,}0308} & \mathbf{0{,}1577} \\ \mathbf{0{,}0924} & \mathbf{0{,}179} & \mathbf{1{,}0404} \end{bmatrix}.$$

Poniższe równania macierzowe możemy zapisać używając przytoczonych już wcześniej symboli:

$$\mathbf{X}-\mathbf{AX}=\mathbf{d}, \quad (\mathbf{I}-\mathbf{A})\mathbf{X}=\mathbf{d}, \quad \mathbf{X}=(\mathbf{I}-\mathbf{A})^{-1},$$

gdzie:

X – macierz produkcji całkowitej (globalnej) informacji w sieci Internet,

A – macierz współczynników informacji,

d – macierz (wektor) popytu użytkowników sieci na informację; wektor ten wskazuje również ważność poszczególnych słów kluczowych względem siebie w tym przypadku.

Cały wielki graf Internetu można traktować jako gospodarkę zamkniętą. W gospodarce zamkniętej całkowita produkcja wytwarzana przez wszystkie gałęzie jest równa jej całkowitej konsumpcji [9]. Prawdziwe jest wówczas równanie $\mathbf{AX} = \mathbf{X}$. Niemożliwe jest jednak zbudowanie układu równań ze wszystkich słów kluczowych, uwzględniających wszystkie języki świata. Dlatego rozpatrujemy wybrany podgraf, znajdujący się w przestrzeni grafu Internetu, składający się z interesujących nas słów. Taki podgraf traktujemy jak gospodarkę otwartą, która oddziałuje z przestrzenią sieci. Dla naszego podgrafu, składającego się ze słów: *praca*, *nauka*, *polityka*, zawsze będzie $\mathbf{AX} \neq \mathbf{X}$, gdyż nie jest on wyizolowany z przestrzeni i samowystarczalny. Współpracuje on z przestrzenią, ponieważ założyliśmy, że każdy podgraf informacyjny wytwarza jeden typ słowa kluczowego lub grupę słów kluczowych i do tego potrzebuje jedno słowo lub grupę słów kluczowych z innego podgrafu informacyjnego, znajdującego się w przestrzeni sieci. Aby więc nasz graf wyprodukował odpowiednio 79 milionów stron internetowych ze słowem *praca*, 42 i 31 milionów stron ze słowem *nauka i polityka*, potrzebuje wsparcia innych słów z przestrzeni. Nasz graf może też udzielać wsparcia innym grafom w przestrzeni produkującym inne słowa, czyli inną informację. Po rozwiązaniu równania postaci: $(\mathbf{I}-\mathbf{A})\mathbf{X}=\mathbf{d}$ otrzymujemy:

$$\begin{bmatrix} \mathbf{1} & \mathbf{-0{,}0343} & \mathbf{-0{,}1827} \\ \mathbf{-0{,}1074} & \mathbf{1} & \mathbf{-0{,}1314} \\ \mathbf{-0{,}0703} & \mathbf{-0{,}169} & \mathbf{1} \end{bmatrix} \begin{bmatrix} \mathbf{79{,}5} \\ \mathbf{42} \\ \mathbf{31{,}2} \end{bmatrix} = \begin{bmatrix} \mathbf{72{,}36} \\ \mathbf{29{,}36} \\ \mathbf{18{,}51} \end{bmatrix} \Rightarrow \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \\ \mathbf{z} \end{bmatrix}.$$

Interpretując rozwiązanie wprost (tak jak w modelu przepływów międzygałęziowych), można byłoby powiedzieć, że w tej konfiguracji słów kluczowych popyt użytkowników na informację wynosi odpowiednio: dla słowa x = 72,36, dla słowa y = 29,36 i dla słowa z = 18,51 milionów jednostek informacyjnych, np. stron www, na których występują te słowa kluczowe w tytule. Trudno zweryfikować ten wynik bez przeprowadzenia badań empirycznych, dotyczących statystyk wyszukiwania informacji przez internautów w obszarze działania wyszukiwarki, z której otrzymaliśmy dane. Aby to zweryfikować, należałoby utworzyć stronę www z analizowanymi słowami *x*, *y*, *z* w tytule i dokonywać obserwacji, poprzez jakie zapytania użytkownicy docierają do naszej strony. Jeżeli faktycznie użytkownicy docierają do strony poprzez zapytania o słowa *x*, *y*, *z* i w takiej proporcji jak w wektorze [**d**], to można powiedzieć, że relacje w grafie informacyjnym są prawidłowe i oddziaływanie informacyjne dla analizowanych węzłów istnieje.

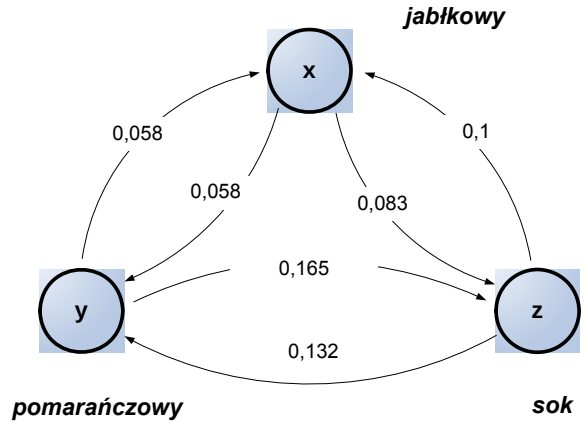
Z pewnością można stwierdzić, że spośród analizowanych wyrazów *x*, *y*, *z* wobec siebie najwyżej w hierarchii jest słowo *x*, później słowo *y*, a na końcu słowo kluczowe *z*. Jeżeli wszystkie elementy wektora [**d**] są dodatnie, można określić procentowy udział jednej jednostki informacji do produkcji innej **AX**/**X** 100%. Inaczej, można powiedzieć, w jakim stopniu słowo kluczowe **x** przyczynia się do produkcji słowa *y*, jako kolejnego węzła w grafie informacji. Jeżeli w wektorze [**d**] pojawi się element ujemny, to znaczy, że analizowany graf informacyjny nie wspiera produkcji tego typu słowa kluczowego i należy zbudować inny graf. Dane słowo kluczowe w tym wypadku musiałoby importować jednostki informacji (linki ze stron www) z przestrzeni, w celu produkcji określonego typu informacji.

5. Przykłady analizy wybranych grafów informacji i zastosowania

Analiza rzeczywistych przypadków grafów informacji, składających się z wybranych słów kluczowych, pozwoli lepiej zrozumieć filozofię i możliwości zastosowania tej metody. Skorzystano z wyszukiwarki google.pl, okresem badania był wrzesień 2008 roku, a dane podano w milionach w celu uproszczenia operacji.

Zapytamy Internet o ważność następujących słów kluczowych względem siebie: *pomidorowy*, *pomarańczowy*, *jabłkowy i sok.* Celem będzie udowodnienie, że im więcej informacji krąży w sieci Internet na temat danego wyrazu, tym jest on ważniejszy względem innych. Także ważność słów kluczowych: *pomarańczowy*, *jabłkowy*, *pomidorowy* względem słowa *sok* pozwoli postawić hipotezę, że najważniejsze słowo po słowie *sok* to produkt najchętniej kupowany przez klientów w rzeczywistym świecie. Najpierw sprawdźmy następujące słowa kluczowe, czyli: *x = jabłkowy*, *y = pomarań-*

czowy, *z = sok.* Później dołożymy czwarte słowo: *q = pomidorowy* i zbudujemy graf informacyjny, składający się już z czterech słów-węzłów. W wyniku przepytania wyszukiwarki internetowej otrzymano graf i wektor globalnej ilości informacji w sieci (dane w mln. stron www).



Rys. 4. Graf zależności informacyjnej dla słów kluczowych: *jabłkowy*, *pomarańczowy*, *sok*

Na podstawie grafu zależności informacji (rys. 4) zbudowano macierz zależności informacji.

		jabłkowy	pomar.	sok
Macierz	jabłkowy	0	0,058	0,083
zależności	pomar.	0,058	0	0,165
informacyjnej	sok	0,1	0,132	0

$$\mathbf{X} = \begin{matrix} 0{,}255 \\ 1{,}86. \\ 192 \end{matrix}$$

Na podstawie wzoru (2) z rozdziału drugiego przekształcamy macierz zależności informacyjnej w macierz współczynników informacji.

$$\mathbf{A} = \begin{bmatrix} \mathbf{0} & \mathbf{0{,}0312} & \mathbf{0{,}0004} \\ \mathbf{0{,}2275} & \mathbf{0} & \mathbf{0{,}0009} \\ \mathbf{0{,}3922} & \mathbf{0{,}071} & \mathbf{0} \end{bmatrix}.$$

Następnie stosujemy znany już wzór $(\mathbf{I} - \mathbf{A})\mathbf{X} = \mathbf{d}$ i w wyniku otrzymujemy wektor popytu końcowego na informację [**d**].

$$(\mathbf{I} - \mathbf{A}) = \begin{matrix} 1 & -0{,}0312 & -0{,}0004 \\ -0{,}2275 & 1 & -0{,}0009 \\ -0{,}3922 & -0{,}071 & 1 \end{matrix} \qquad [\mathbf{d}] = \begin{matrix} 0{,}114 & x \\ 1{,}637 & \Rightarrow y \\ 191{,}768 & z \end{matrix}$$

Na podstawie wartości elementów wektora [**d**] stwierdzamy, że najważniejszym słowem kluczowym w tym grafie jest *z*, tj. słowo *sok*, następnie słowo *pomarańczowy*, a później słowo *jabłkowy*.

$$\mathbf{AX/X}\ 100\% = \begin{matrix} 55{,}29\% \\ 11{,}99\% \\ 0{,}12\% \end{matrix} \begin{matrix} x \\ \Rightarrow y. \\ z \end{matrix}$$

Można też stwierdzić, że słowa kluczowe *jabłkowy* i *pomarańczowy* wpływają w 0,12% na produkcję słowa *sok,* które samo w sobie jest silnym węzłem, wzmacniającym pozycję innych słów w rozpatrywanym grafie. Jest to stosunkowo mały udział, więc w sieci na produkcję witryn ze słowem kluczowym *sok* w nagłówku muszą mieć wpływ inne słowa z przestrzeni. Za to za produkcję informacji ze słowem kluczowym *jabłkowy* w 55,29% odpowiadają słowa *sok* i *pomarańczowy* w tej konfiguracji.

Teraz budujemy graf składający się z czterech węzłów. Do analizy włączamy czwarte słowo kluczowe: *q* = *pomidorowy*. Przepytujemy wyszukiwarkę i sprawdzamy relacje. W wyniku otrzymujemy następujące macierze danych.

	sok	pomar.	jabłkowy	pomidor.	**(I – A)**			
sok	0	0,145	0,1	0,12	1	–0,0718	–0,3876	–0,4781
pomarańczowy	0,166	0	0,058	0,038	–0,0008	1	–0,2248	–0,1514
jabłkowy	0,08	0,058	0	0,035	–0,0004	–0,0287	1	–0,1394
pomidorowy	0,12	0,038	0,086	0	–0,0006	–0,0188	–0,3333	1

A

0	0,07178	0,3876	0,47809
0,00081	0	0,22481	0,15139
0,00039	0,02871	0	0,13944
0,00059	0,01881	0,33333	0

X =

204	sok
2,02	pomar.
0,258	jabłkowy
0,251	pomidor.

Stosujemy wzór (**I** – **A**)**X** = **d** i otrzymujemy wektor popytu końcowego na informację [**d**].

	[**d**]	**AX/X**	**Udział w rynku**
sok	203,635	0,18%	–
pomarańczowy	1,758	12,97%	największy
jabłkowy	0,085	67,05%	średni
pomidorowy	0,007	97,21%	najmniejszy

Na podstawie wartości elementów wektora [**d**] stwierdzamy, że najważniejszym słowem kluczowym w tym grafie jest słowo *sok*, następnie słowo *pomarańczowy*, a później słowo *jabłkowy*, na końcu wyraz *pomidorowy*. Widzimy również procentowy

udział (wpływ) rozpatrywanych słów kluczowych na produkcję pozostałych słów kluczowych w tym grafie. Na produkcję stron internetowych ze słowem kluczowym *pomidorowy* w 97,21% wpływają słowa: *sok*, *pomarańczowy* i *jabłkowy*. Natomiast na wytwarzanie w sieci słowa *sok* pozostałe rozpatrywane słowa wpływają jedynie w 0,18%. Wynika z tego, że słabe słowa kluczowe wchodzą w interakcję z mocnymi, tak jak słowo *pomidorowy*, którego istnienie w dużym stopniu uzależnione jest od słowa *sok*. Słabe słowa kluczowe podnoszą tym samym swoją pozycję w hierarchii ważności informacji sieci. Tak jak słowa *pomarańczowy*, *jabłkowy* i *pomidorowy* powstają w odpowiednich proporcjach dzięki słowu kluczowemu *sok*. Węzeł ten jest mocny w tej konfiguracji słów i wzmacnia pozycję pozostałych w rozpatrywanym grafie.

Trudno oprzeć się pokusie porównania wyników, pochodzących z wirtualnego świata informacji, do realiów świata rzeczywistego. Czy zależności dotyczące rzeczywistego spożycia soku pomarańczowego, jabłkowego i pomidorowego są w jakiś sposób proporcjonalne do ilości informacji krążących w sieci Internetu na ten temat? Wyobraźmy sobie sytuację, że mamy do dyspozycji tylko trzy wymienione smaki soków: względem słowa *sok* kolejne wartości otrzymuje więc słowo kluczowe *pomarańczowy*, *jabłkowy* i *pomidorowy*. Jeżeli spojrzeć na wartości elementów wektora

$$[\mathbf{d}] = \begin{bmatrix} 203{,}635 \\ 1{,}758 \\ 0{,}085 \\ 0{,}007 \end{bmatrix} \Rightarrow \begin{bmatrix} \text{sok} \\ \text{pomarańczowy} \\ \text{jabłkowy} \\ \text{pomidorowy} \end{bmatrix},$$

to można przypuszczać, że największe spożycie jest dla soku pomarańczowego, później jabłkowego, a na końcu pomidorowego. Celem niniejszego opracowania nie jest precyzyjne wskazanie wyniku, ale ogólne pokazanie istnienia takiej zależności. Aby odnieść wartość wektora [**d**] do rzeczywistości, należałoby znormalizować wyniki według przyjaznej dla odbiorcy skali i wziąć pod uwagę odpowiednią ilość smaków soków występujących na rynku oraz w obszarze działania wyszukiwarki. Więcej informacji na ten temat można znaleźć w innych publikacjach autora.

Badania firmy Nielsen ze stycznia 2008 roku pokazują, że w kategorii soków naturalnych największym uznaniem wśród polskich konsumentów cieszą się dwa smaki: pomarańczowy (34-procentowy ilościowy udział w rynku) oraz jabłkowy (27,2-procentowy). Lubimy także soki: pomidorowe (15,9-procentowy ilościowy udział w sprzedaży), a także soki warzywne (4-procentowy), grapefruitowe (3,7-procentowy) i wielowocowe (3,6-procentowy). [*Źródło: http://www.portalspozywczy.pl*]. Dane te potwierdzają, iż częściej konsumenci sięgają po soki, wokół których krąży więcej informacji w sieci, i których słowa kluczowe są ważniejsze w hierarchii informacji względem innych analizowanych w grafie informacyjnym.

Dokonując porównań słów kluczowych w macierzy zależności informacyjnej można dojść do wniosku, iż zawsze są one diagonalne. Dzieje się tak dlatego, że graf

informacyjny nie zużywa własnej informacji do produkcji następnej. Może on wspierać przestrzeń sieci w budowaniu informacji lub pobierać z niej słowa kluczowe do powiększania i budowy własnych zasobów informacyjnych. Ciekawym zjawiskiem jest również to, iż słowa bliskoznaczne (synonimy) przeważnie tworzą macierze diagonalne i symetryczne.

6. Wnioski

Podane przykłady świadczą o tym, iż możliwe jest zastosowanie modelu przepływów międzygałęziowych do określania zapotrzebowania na informację w globalnej sieci oraz wyznaczania hierarchii ważności tejże. Metoda pozwala określać ekspansję wielkiego grafu Internetu, który rozwija się w kierunku informacji, na którą jest największe zapotrzebowanie. Przed napisaniem niniejszego artykułu autor przeprowadził wiele podobnych analiz dla różnych sekwencji słów kluczowych. Sprawdzono na przykład relacje, jakie zachodzą m.in. wśród słów kluczowych największych miast w Polsce. Wyniki badań potwierdzają, iż konkurencyjność regionu lub miasta ma swoje odzwierciedlenie w postaci ilości produkcji informacji w sieci. Wyniki tych wstępnych badań otwierają pole do dyskusji na temat roli i znaczenia informacji w sieci jako narzędzia prognostycznego. Potwierdzają hipotezę o dużej sile ciążenia wielkich grafów informacji, które przyciągają mniejsze podgrafy, niosące mniej ważną informację.

Dziedziny życia o kluczowym dla ludzi znaczeniu również mają swoją reprezentację informacyjną w Internecie. Analizując grafy zależności informacyjnej dochodzimy do wniosku, iż potwierdza się hierarchia ważności potrzeb Maslowa. Pierwsze miejsce zajmują takie elementy jak: życie, zdrowie, bezpieczeństwo, rodzina, przyjaciele, religia (wiara, za którą niektórzy są skłonni oddać życie). Na drugim znalazły się kariera, edukacja oraz styl życia. Na trzecim: polityczne i społeczne prawdy (poglądy), na czwartym zaś filozoficzne wierzenia, idee i myśli. Należy zaznaczyć, iż zarówno grupy, korporacje, jak i państwa reprezentują podobny system wartości [1]. Wreszcie grafy i macierze zależności informacyjnej mogą być źródłem danych dla innych metod wspomagających podejmowanie decyzji wielokryterialnych, np. AHP/ANP [8].

Następnym aspektem, na który warto zwrócić uwagę to pozycjonowanie stron www z użyciem tej metody. Wyszukiwarki internetowe nie dorównują jeszcze człowiekowi, ale są coraz bardziej inteligentne. Nie tylko rozpoznają słowa kluczowe w nagłówku czy opisie, lecz analizują także treści strony i oddziaływania zachodzące między słowami kluczowymi (węzłami). W określaniu ważności danej strony, a tym samym jej pozycji w sieci www, decyduje już nie tylko ilość, ale jakość linków, jakie

do niej prowadzą. Można zaobserwować zjawisko, iż wyżej pozycjonowane są witryny posiadające linki kontekstowe, pochodzące z grafów informacji zawierających informację bliskoznaczną czy podobną do pozycjonowanej strony. Witryny posiadające linki przypadkowe, pochodzące od zupełnie innej treści niż prezentowana na pozycjonowanej stronie, mogą w przyszłości w ogóle nie liczyć się w sieci. Znajdą się wówczas na odległych miejscach w wyszukiwarkach, a ich dostępność informacyjna będzie bardzo mała. Analiza grafów zależności informacyjnej pozwoli optymalnie dobrać słowa kluczowe na naszym serwisie i zwiększyć pozycję strony w wyszukiwarce. Autor zastosował tę metodę do pozycjonowania stron w wyszukiwarkach z bardzo dobrym rezultatem. Na podstawie analizy grafu informacji w treści strony www umieszczono słowa kluczowe, mające wsparcie silnych grafów informacyjnych. Wynik jest taki, iż w odpowiedzi na zapytanie o którekolwiek ze słów kluczy wyszukiwarka google.pl wyrzuca naszą stronę zawsze w pierwszej dziesiątce wyników (na kilka milionów odpowiedzi). Macierze i grafy zależności informacyjnej mogą więc pomóc w doborze słów kluczowych i nawet całej treści witryny, w celu jej lepszego pozycjonowania i maksymalnej dostępności informacyjnej dla przeciętnych użytkowników Internetu.

Bibliografia

[1] ADAMUS W., GRĘDA A., *Wspomaganie decyzji wielokryterialnych w rozwiązywaniu wybranych problemów organizacyjnych i menedżerskich*, Badania Operacyjne i Decyzje, 2005, nr 2.
[2] CHIANG A.G., *Podstawy ekonomii matematycznej*, Państwowe Wydawnictwo Ekonomiczne, Warszawa 1994.
[3] CZERWIŃSKI Z., *Matematyka na usługach ekonomii*, PWN, Warszawa 1980.
[4] DOWLING E.T., *Introduction to Mathematical Economics*, McGraw-Hill Professional, 2000.
[5] KACPRZAK D., *Analiza modelu Leontiewa z użyciem skierowanych liczb rozmytych – S/WI/1/07*, Wydział Informatyki, Politechnika Białostocka, Białystok 2007.
[6] KIFNER T., *Polityka bezpieczeństwa i ochrony informacji*, Helion, Gliwice 1999.
[7] KOWALCZYK E., *O istocie informacji*, WKŁ, Warszawa 1981.
[8] SAATY T.L., *Decision Making – The Analytic Hierarchy and Network Processes (AHP/ANP)*, Journal of Systems Science and Systems Engineering, published at Tsinghua University, Beijing, 2004b, Vol. 13, No. 1, 1–34, March.
[9] http://math.fullerton.edu/mathews/n2003/LeontiefModelMod.html.

Use of the input–output matrix to determine the hierarchy of information on the Internet

The author examines the Internet network as a global economy in which the only output is information. It is assumed that there is a certain hierarchy of keywords in the net space. Important keywords

create large information graphs, which attract smaller graphs consisting of less important data. Information graphs are linked to each other like branches of production in classical economics. Thus, due to the huge amount of information on the net, it is difficult to measure its quality, or whether it is true or false. A non-standard method of measuring information is proposed, as well as analysis of the weights between each of the keywords. Search results from Internet search engines were used. The input–output matrix was used to analyze the amount of information, as well as dependencies which occur between individual keywords.

Keywords: *Internet network, production of information, vector, graph of information, input–output model, keyword analysis*