Przemysław Kazienko: *Strukturalne podobieństwo dokumentów hipertekstowych*. Informacja Wiedza Gospodarka. (VI Krajowe Forum Informacji Naukowej i Technicznej, Katowice, 2001). Prace PTIN NR 4, Polskie Towarzystwo Informacji Naukowej, Warszawa 2001, s.244-253.

Przemysław **KAZIENKO**

Strukturalne podobieństwo dokumentów hipertekstowych

Cechą charakterystyczną dokumentów hipertekstowych są odsyłacze, które tworzą strukturę systemu hipertekstowego. Zakładając, że odsyłacze niosą ze sobą informację o powiązaniach semantycznych między dokumentami, zaproponowano wykorzystać elementy struktury do wyznaczenia podobieństwa pomiędzy dokumentami hipertekstowymi. W pracy przedstawiono cztery, nowe funkcje podobieństwa strukturalnego, które mogą okazać się szczególnie przydatne w hipermedialnym systemie WWW. W funkcjach tych wykorzystano możliwość automatycznego określenia rodzaju odsyłaczy i podziału na odsyłacze semantyczne i nawigacyjne. Zaprezentowano także obszary zastosowań funkcji podobieństwa strukturalnego, szczególnie przy wyszukiwaniu informacji w sieci WWW.

Links are the specific feature of hypertext documents and the primary part of hypertext structure. The important information about semantic relations between documents is encapsulated within them. This information can be used to point out similarity function.
In the paper four, new structure similarity functions were described. They can be usefull in the WWW environment. The automatic link type detection was utilized in these functions (semantic and navigational links can be distinguished). Similarity function application areas were presented as well.

1. WPROWADZENIE

Elementem charakterystycznym dokumentów hipertekstowych, odróżniającym go od innych rodzajów dokumentów są odsyłacze. Tworzą one strukturę systemu hipertekstowego. Łącząc dokumenty, równocześnie niosą ze sobą informacje o związkach semantycznych pomiędzy dokumentami.

Garzotto, Paolini i Schwabe [6] twierdzą, że odsyłacze pełnią w systemie hipertekstowym dwie podstawowe role: reprezentacyjną (ujmując i prezentując relacje między porcjami informacji w tekście) oraz nawigacyjną (obejmując ścieżki poruszania się użytkownika po systemie).

Nie rzadko te dwie role przenikają się nawzajem, czasem są rozdzielone. Ten drugi przypadek występuje w sytuacjach, w których istniejące związki miedzy poszczególnymi rodzajami informacji, umieszczonymi w różnych dokumentach, są nieodpowiednie dla konkretnej koncepcji nawigacji po systemie ustalonej przez autorów, to znaczy, odsyłacze służące wyłącznie poruszaniu się, mogą łączyć ze sobą dokumenty o słabszych związkach semantycznych.

Wielu autorów [1, 5, 8, 9] biorąc pod uwagę dwie, wymienione zasadnicze role (reprezentacyjną i nawigacyjną), jakie odgrywają odsyłacze w systemach hipertekstowych, rozróżnia dwa główne typy odsyłaczy:

- **Semantyczne** (znaczeniowe), czyli bazujące na treści, mające za zadanie wiązać dokumenty mieszczące się w tej samej lub pokrewnej tematyce.
- **Organizacyjne**, czyli odsyłające, których głównym celem jest lepsza (szybsza, łatwiejsza) nawigacja po systemie. W związku z tym, bywają one także nazywane **nawigacyjnymi** [9].

Na podstawie badań Haas i Grams [7] można stwierdzić, że dla hipertekstowego środowiska WWW z dużym prawdopodobieństwem (ponad 80%) można automatycznie rozróżnić odsyłacze nawigacyjne od semantycznych.

2. ELEMENTY SYSTEMU HIPERTEKSTOWEGO

Skończony zbiór wszystkich dokumentów hipertekstowych D^E, wraz ze zbiorem L^E wszystkich odsyłaczy wychodzących z dokumentów, należących do D^E tworzą parę (D^E,L^E), którą nazwijmy ***środowiskiem hipertekstowym***. Zauważmy, przy tym, że w środowisku otwartym (takim jest np. zbiór stron WWW) odsyłacze ze zbioru L^E wychodzą zawsze z dokumentu $d_i \in D^E$, jednak prowadzą one do dokumentu d_j ze zbioru $D^E \cup D^-$, gdzie D^- to zbiór dokumentów nie należących do D^E, reprezentujących dokumenty jeszcze lub już nieistniejące, tzn. odsyłacze mogą prowadzić do dokumentów nieistniejących. W tak określonym środowisku hipertekstowym można wydzielić ***kolekcję hipertekstową***, będącą parą (D,L), taką że $D \subseteq D^E$, zaś $L \subseteq L^E$ jest zbiorem wszystkich odsyłaczy ze zbioru L^E, dla których dokument początkowy należy do D.

3. ZAŁOŻENIA DLA FUNKCJI PODOBIEŃSTWA STRUKTURALNEGO

Traktując odsyłacze jako główne źródło informacji o podobieństwie dokumentów, można opracować funkcję podobieństwa strukturalnego (odsyłacze tworzą strukturę systemu hipertekstowego). Przyjmijmy pewne założenia przy wyznaczaniu funkcji podobieństwa strukturalnego. W funkcji tej będą uwzględniane:

a) Liczba odsyłaczy łączących oba dokumenty (odsyłacze bezpośrednie).
b) Liczba wszystkich odsyłaczy, które wychodzą z obu dokumentów. Pozwala to na ważenie odsyłaczy łączących dokumenty. Waga jest większa wtedy, gdy są to je-

dyne odsyłacze występujące w tych dokumentach; mniejsza - gdy są one jednymi z wielu.

c) Liczba i rodzaj bezpośrednich wspólnych potomków i przodków (w ramach kolekcji i poza nią).

d) Rodzaj odsyłaczy według podziału na semantyczne i nawigacyjne przyjmując, że odsyłacze semantyczne lepiej niż nawigacyjne odzwierciedlają podobieństwo tematyczne między dokumentami.

Załóżmy także, że zbiorem wartości funkcji podobieństwa jest przedział [0,1].

4. PIERWOTNE PODOBIEŃSTWO STRUKTURALNE

Poniżej zostały przedstawione cztery wersje funkcji podobieństwa, które nazwijmy ***pierwotnym podobieństwem strukturalnym***. Oznaczono je kolejnymi, wielkimi literami alfabetu, tzn. A, B, C oraz D. Wszystkie (prócz wersji D) bazują na odsyłaczach łączących dokumenty w ramach kolekcji hipertekstowej (w ramach zbioru D), czyli tych, dla których dokument początkowy i końcowy należy do zbioru D.

Chcąc zwiększyć przejrzystość zapisu, oznaczmy funkcję pierwotnego podobieństwa strukturalnego między dwoma dokumentami $d_i,d_j \in D$, obliczoną wg wersji A, jako $S_A^{ps}(d_i,d_j)$, wg wersji B – jako $S_B^{ps}(d_i,d_j)$, wg wersji C – $S_C^{ps}(d_i,d_j)$, a wg wersji D – $S_D^{ps}(d_i,d_j)$.

4.1. WERSJA A PIERWOTNEGO PODOBIEŃSTWA STRUKTURALNEGO

Funkcja podobieństwa $S_A^{ps}(d_i,d_j)$ opiera się na znormalizowanej, średniej arytmetycznej liczby odsyłaczy prowadzących z jednego dokumentu do drugiego, dla których obliczamy wartość pierwotnego podobieństwa strukturalnego. Normalizacja dokonuje się przez podzielenie przez liczbę wszystkich odsyłaczy wychodzących z obu dokumentów do innych dokumentów w ramach kolekcji D:

$$S_A^{ps}(d_i,d_j)=\begin{cases}\frac{1}{2}\left(\frac{nl_{ij}}{nl_i}+\frac{nl_{ji}}{nl_j}\right), & nl_i>0, nl_j>0\\ \frac{1}{2}\frac{nl_{ji}}{nl_j}, & nl_i=0, nl_j>0\\ \frac{1}{2}\frac{nl_{ij}}{nl_i}, & nl_i>0, nl_j=0\\ 0, & nl_i=0, nl_j=0\end{cases}, \qquad (1)$$

gdzie: nl_{ij} – względna liczba odsyłaczy z d_i do d_j; nl_i – liczba wszystkich odsyłaczy wychodzących z dokumentu d_i, a prowadzących do innych dokumentów w ramach danej kolekcji D:

Funkcję $S_A^{ps}(d_i,d_j)$ można traktować jako średnią arytmetyczną podobieństw między d_i i d_j oraz w drugą stronę: między d_j i d_i. Jest to średnia arytmetyczna funkcji Chena [3], liczona w obie strony. Dzięki temu uzyskuje się symetrię funkcji. Wartości $S_A^{ps}(d_i,d_j)$ zawierają się w przedziale [0,1], ponieważ $nl_i \geq nl_{ij} \geq 0$, a $nl_j \geq nl_{ji} \geq 0$.

Wadą wersji A jest to, że w przypadku, gdy nie istnieją odsyłacze w drugą stronę (tzn. $nl_{ij}=0$ i $nl_{ji}>0$ lub $nl_{ij}>0$ i $nl_{ji}=0$), a sytuacja taka bardzo często się zdarza w systemach hipertekstowych (zwłaszcza w przypadku stron WWW), wtedy jeden z ułamków w pierwszym wierszu ma wartość 0, zaś drugi jest dzielony przez 2. W efekcie, jeżeli istnieje odsyłacz między dokumentem d_i oraz d_j (a nawet wtedy, gdy jest ich dużo), lecz nie ma odsyłacza w drugą stronę, to i tak wartość pierwotnego podobieństwa strukturalnego nie może przekroczyć ½.

Względna liczba odsyłaczy nl_{ij} określa liczbę odsyłaczy łączących dwa dokumenty, ale z uwzględnieniem ich rodzaju, tzn. odsyłacze semantyczne mogą być ważniejsze od odsyłaczy nawigacyjnych, czyli:

$$nl_{ij} = nl_{ij}^s + \lambda\, nl_{ij}^n, \qquad (2)$$

gdzie: nl_{ij}^s – liczba odsyłaczy semantycznych z d_i do d_j; nl_{ij}^n – liczba odsyłaczy nawigacyjnych z d_i do d_j; λ – współczynnik określający znaczenie odsyłaczy nawigacyjnych, $\lambda \in [0,1]$.

Dzięki λ możliwe jest regulowanie znaczenia odsyłaczy nawigacyjnych. W przypadku $\lambda=1$ tracimy zupełnie rozróżnienie na odsyłacze nawigacyjne i semantyczne. Dla $\lambda=0$ odsyłacze nawigacyjne są pomijane. Wydaje się, że współczynnik λ powinien być raczej bliższy jeden niż zero. Przemawia za tym spostrzeżenie, że odsyłacze nawigacyjne także niosą ze sobą informację o powiązaniach tematycznych między dokumentami (oczywiście waga takiej informacji jest mniejsza niż w przypadku odsyłaczy semantycznych). Wynika to z tego, że autor poprzez umieszczenie odsyłacza nawigacyjnego przewiduje, iż użytkownik zainteresowany danym dokumentem rów-

nież będzie zainteresowany dokumentem, do którego prowadzi ten odsyłacz [13]. Oba dokumenty są więc podobne z punktu widzenia potrzeb informacyjnych

Po uwzględnieniu (2) we wzorze (1), otrzymujemy:

$$S_A^{ps}(d_i,d_j)=\begin{cases}\frac{1}{2}\left(\frac{nl_{ij}^s+\lambda\, nl_{ij}^n}{nl_i}+\frac{nl_{ji}^s+\lambda\, nl_{ji}^n}{nl_j}\right), & nl_i>0, nl_j>0\\ \frac{1}{2}\frac{nl_{ji}^s+\lambda\, nl_{ji}^n}{nl_j}, & nl_i=0, nl_j>0\\ \frac{1}{2}\frac{nl_{ij}^s+\lambda\, nl_{ij}^n}{nl_i}, & nl_i>0, nl_j=0\\ 0, & nl_i=0, nl_j=0\end{cases}.$$

Jeżeli $nl_{ij}^s=nl_{ji}^s=0$, czyli gdy odsyłacze łączące d_i oraz d_j ze sobą są odsyłaczami nawigacyjnymi, wtedy $S_A^{ps}(d_i,d_j)$ ma wartość λ razy mniejszą niż wtedy, gdy wszystkie odsyłacze są semantyczne.

4.2. WERSJA B PIERWOTNEGO PODOBIEŃSTWA STRUKTURALNEGO

Funkcja $S_B^{ps}(d_i,d_j)$ jest zmodyfikowaną postacią funkcji $S_A^{ps}(d_i,d_j)$:

$$S_B^{ps}(d_i,d_j)=\begin{cases}\frac{nl_{ij}+nl_{ji}}{nl_i+nl_j}, & \text{dla } nl_i+nl_j>0\\ 0, & \text{dla } nl_i+nl_j=0\end{cases}.$$

Podobnie jak w przypadku $S_A^{ps}(d_i,d_j)$, zbiór wartości funkcji $S_B^{ps}(d_i,d_j)$ to przedział [0,1].

Dzięki wyeliminowaniu współczynnika ½ przed ułamkiem uzyskuje się to, że pierwotne podobieństwo strukturalne może nawet osiągnąć wartość jeden, mimo tego, że nie będą istnieć odsyłacze w obie strony między dokumentem d_i oraz d_j.

Uwzględniając mniejsze znaczenie odsyłaczy nawigacyjnych w stosunku do semantycznych, czyli stosując wzór (2) do powyższego, otrzymujemy:

$$S_B^{ps}(d_i,d_j)=\begin{cases}\frac{\left(nl_{ij}^s+\lambda\, nl_{ij}^n\right)+\left(nl_{ji}^s+\lambda\, nl_{ji}^n\right)}{nl_i+nl_j}, & \text{dla } nl_i+nl_j>0\\ 0, & \text{dla } nl_i+nl_j=0\end{cases}.$$

4.3. WERSJA C PIERWOTNEGO PODOBIEŃSTWA STRUKTURALNEGO

W wersji A oraz B funkcji pierwotnego podobieństwa strukturalnego nie były uwzględnione gęstości odsyłaczy. Jeżeli między dokumentem d_i oraz d_j istnieje jeden odsyłacz semantyczny (nl_{ij}=1) i w drugą stronę także jest jeden odsyłacz semantyczny (nl_{ji}=1), oraz z każdego z tych dokumentów nie wychodzą inne odsyłacze

(nl_i=nl_j=1), wtedy $S_A^{ps}(d_i,d_j)$= $S_B^{ps}(d_i,d_j)$=1. Dokumenty są maksymalnie podobne do siebie. Tyle samo wyniosą wartości pierwotnego podobieństwa strukturalnego dla obu wersji A i B (tzn. 1) w sytuacji, gdy między dokumentami jest po pięć odsyłaczy (nl_{ij}=nl_{ji}=5) i ponownie nie wychodzą z nich inne odsyłacze (nl_i=nl_j=5).

Gęstość odsyłaczy uwzględniono w definicji funkcji $S_C^{ps}(d_i,d_j)$:

$$S_C^{sp}(d_i,d_j)=\begin{cases}\dfrac{nl_{ij}+nl_{ji}}{nl_{max}}, & nl_{max}>0\\ 0, & nl_{max}=0\end{cases},$$

gdzie: nl_{max} – maksymalna liczba odsyłaczy łączących dwa dokumenty, należące do kolekcji D, tzn. $nl_{max} = max\{nl_{ij}+nl_{ji}: i\neq j, d_i,d_j\in D\}$.

Normalizacja (mianownik ułamka) następuje tutaj nie względem liczby odsyłaczy wychodzących z dokumentów – między którymi liczymy podobieństwo – lecz względem maksymalnej liczby odsyłaczy łączących jakiekolwiek dwa dokumenty w kolekcji, czyli nl_{max}.

Zbiór wartości funkcji $S_C^{ps}(d_i,d_j)$ to przedział [0,1], ponieważ $nl_{max}\geq nl_{ij}+nl_{ji}$.

Jeżeli nl_{max}=0, to w kolekcji D nie występują odsyłacze między dokumentami i kolekcja ta w ogóle nie posiada charakteru hipertekstowego.

Stosując wzór (2) do powyższego otrzymujemy:

$$S_C^{ps}(d_i,d_j)=\begin{cases}\dfrac{\left(nl_{ij}^s+\lambda\, nl_{ij}^n\right)+\left(nl_{ji}^s+\lambda\, nl_{ji}^n\right)}{nl_{max}}, & nl_{max}>0\\ 0, & nl_{max}=0\end{cases}.$$

4.4. WERSJA D PIERWOTNEGO PODOBIEŃSTWA STRUKTURALNEGO

Poprzednie wersje postaci funkcji pierwotnego podobieństwa strukturalnego, tj. A, B oraz C uwzględniały jedynie (w liczniku) odsyłacze łączące ze sobą bezpośrednio dokument d_i oraz d_j. Jednakże – już wiele lat temu – przy opracowywaniu różnych funkcji podobieństwa, związanych ze wzajemnym cytowaniem w tekstach naukowych, dostrzeżono, że dokumenty, które cytują te same, inne dokumenty nawet, jeżeli nie cytują siebie nawzajem, są do siebie podobne. Można założyć, że podobna hipoteza jest słuszna dla środowiska hipertekstowego. Rozróżnijmy dwa przypadki:

1. Ponieważ dwa dokumenty d_i oraz d_j wskazują na ten sam inny dokument (cytują go) d_k, więc d_i oraz d_j są do siebie podobne (wspólni potomkowie - dzieci).

2. Dokument d_k wskazuje zarówno na dokument d_i jak i d_j, więc dokumenty d_i oraz d_j są do siebie podobne (wspólni przodkowie - rodzice); dokumenty są współcytowane.

W pierwszym punkcie uwzględniane są odsyłacze, które nazwijmy ***pośrednimi odsyłaczami cytowania***. W drugim występują ***pośrednie odsyłacze współcytowania***.

Na rys. 1 przedstawiona jest kolekcja dokumentów tekstowych (zbiór D) wraz z odsyłaczami (zbiór L), czyli kolekcja hipertekstowa (D,L). Przykładem pierwszego przypadku na tym rysunku jest dokument 1 (d_i) oraz 6 (d_j), które wskazują na ten sam dokument 2 (d_k). Może zaistnieć również taka sytuacja, w której dokument d_k nie należy do kolekcji, czyli $d_k \in (D^E \cup D^-) \backslash D$. Na przykład, odsyłacze z dokumentu 10 oraz 11 prowadzą do dokumentu 22, leżącego poza zbiorem D grupowanej kolekcji. Drugi przypadek to 8 (d_k), z którego prowadzą odsyłacze zarówno do dokumentu 4 (d_i), jak i 9 (d_j). Tutaj także może zaistnieć sytuacja, w której dokument wskazujący na te, między którymi obliczane jest podobieństwo, czyli d_k, leży poza zbiorem D, jednak nie może to być element zbioru D^-. Przykładem tego jest dokument 23, nie należący do kolekcji D, posiadający odsyłacze do dokumentu 6 oraz 10.

Intuicyjnie można domniemywać, że podobieństwo wynikające tylko z istnienia pośrednich odsyłaczy cytowania i współcytowania powinno mieć mniejsze znaczenie (wagę) niż podobieństwo, które jest związane z odsyłaczami bezpośrednio łączącymi dwa dokumenty (cytowanie siebie nawzajem). W związku z tym, proponuje się wprowadzenie współczynnika ważności pośrednich odsyłaczy cytowania i współcytowania μ ($\mu \in [0,1]$), zmniejszającego ich wagę (znaczenie) w stosunku do odsyłaczy bezpośrednio łączących dwa dokumenty.

Przy opracowywaniu funkcji pierwotnego podobieństwa strukturalnego $S'^{ps}_D(d_i,d_j)$ oparto się na wersji B. Jako element normalizacji liczby odsyłaczy pośrednich cytowania i bezpośrednich łącznie, przyjęto liczbę wszystkich odsyłaczy wychodzących z dokumentów d_i oraz d_j, czyli odpowiednio $nl_i^{\rightarrow}$ i $nl_j^{\rightarrow}$. Różnica między nl_i i nl_j, które występują w wersji B, a $nl_i^{\rightarrow}$ i $nl_j^{\rightarrow}$ polega na tym, że te ostatnie uwzględniają także odsyłacze prowadzące do dokumentów spoza kolekcji (zbiór dokumentów końcowych tych odsyłaczy to $D^E \cup D^-$ a nie tylko D). Do normalizacji pośrednich odsyła-

czy współcytowania wykorzystano liczbę wszystkich odsyłaczy wychodzących ze wspólnych dla d_i i d_j przodków-rodziców (zbiór o liczności $nl_{ij}^{\Leftrightarrow}$). W efekcie powstał wzór:

$$S_D^{ps}(d_i,d_j)=\begin{cases}\dfrac{nl_{ij}+nl_{ji}+\mu\left(nl_{ij}^{\Rightarrow}+nl_{ji}^{\Rightarrow}+nl_{ij}^{\Leftarrow}+nl_{ji}^{\Leftarrow}\right)}{nl_i^{\rightarrow}+nl_j^{\rightarrow}+nl_{ij}^{\Leftrightarrow}}, & \text{dla } nl_i^{\rightarrow}+nl_j^{\rightarrow}+nl_{ij}^{\Leftrightarrow}>0\\ 0, & \text{dla } nl_i^{\rightarrow}+nl_j^{\rightarrow}+nl_{ij}^{\Leftrightarrow}=0\end{cases},$$

(3)

gdzie: $nl_{ij}^{\Rightarrow}$ – względna liczba wszystkich, pośrednich odsyłaczy cytowania wychodzących z dokumentu d_i, względem dokumentu d_j; $nl_{ij}^{\Leftarrow}$ – względna liczba wszystkich pośrednich odsyłaczy współcytowania prowadzących do dokumentu d_i, względem dokumentu d_j.

Względne liczby odsyłaczy cytowania i współcytowania można wyrazić za pomocą wzorów analogicznych do (2):

$$nl_{ij}^{\Rightarrow}=nl_{ij}^{s\Rightarrow}+\lambda\, nl_{ij}^{n\Rightarrow},\qquad nl_{ij}^{\Leftarrow}=nl_{ij}^{s\Leftarrow}+\lambda\, nl_{ij}^{n\Leftarrow},\qquad (4)$$

gdzie: $nl_{ij}^{s\Rightarrow}$, $nl_{ij}^{n\Rightarrow}$ – liczba odpowiednio semantycznych, nawigacyjnych, pośrednich odsyłaczy cytowania z dokumentu d_i, względem dokumentu d_j; $nl_{ij}^{s\Leftarrow}$, $nl_{ij}^{n\Leftarrow}$ – liczba odpowiednio semantycznych, nawigacyjnych, pośrednich odsyłaczy współcytowania dokumentu d_i, względem dokumentu d_j.

Stosując do (3) wzory (2) oraz (4), otrzymujemy:

$$S_D^{ps}(d_i,d_j)=\begin{cases}\dfrac{nl_{ij}^{s}+nl_{ji}^{s}+\lambda\left(nl_{ij}^{n}+nl_{ji}^{n}\right)+\mu\left(nl_{ij}^{s\Rightarrow}+nl_{ji}^{s\Rightarrow}+nl_{ij}^{s\Leftarrow}+nl_{ji}^{s\Leftarrow}\right)}{nl_i^{\rightarrow}+nl_j^{\rightarrow}+nl_{ij}^{\Leftrightarrow}}+\\ \dfrac{+\mu\,\lambda\left(nl_{ij}^{n\Rightarrow}+nl_{ji}^{n\Rightarrow}+nl_{ij}^{n\Leftarrow}+nl_{ji}^{n\Leftarrow}\right)}{}, & \text{dla } nl_i^{\rightarrow}+nl_j^{\rightarrow}+nl_{ij}^{\Leftrightarrow}>0\\ 0, & \text{dla } nl_i^{\rightarrow}+nl_j^{\rightarrow}+nl_{ij}^{\Leftrightarrow}=0\end{cases}$$

4.5. MODYFIKACJE WERSJI D PIERWOTNEGO PODOBIEŃSTWA STRUKTURALNEGO

W większości środowisk hipertekstowych, w tym także w systemie WWW, pełna informacja o odsyłaczach jest umieszczana w treści dokumentu. Mając dostęp do pełnej treści posiada się więc także możliwość wykorzystania informacji semantycznej, którą niosą ze sobą odsyłacze. W związku z tym ustalenie wspólnych potomków-dzieci, poprzez porównanie adresów węzłów docelowych odsyłaczy nie nastrę-

cza większych problemów. Aby jednak wyznaczyć wspólnych przodków-rodziców i obliczyć liczby odpowiednich odsyłaczy z nich wychodzących ($nl_{ij}^{\Leftarrow}$, $nl_{ji}^{\Leftarrow}$, $nl_{ij}^{\Leftrightarrow}$), należy mieć dostęp do treści owych przodków. Jest to możliwe wtedy, gdy grupowaną kolekcją jest pewien zamknięty system hipertekstowy, np. system autorski lub pojedynczy serwis informacyjny WWW.

Poważny problem pojawia się jednak wtedy, gdy kolekcją hipertekstową (D,L) będzie część większego środowiska, np. zbiór stron WWW zwróconych przez wyszukiwarki internetowe (część całego systemu WWW). Jak uzyskać w takim przypadku informację o treści przodków? Wymaga to przecież ustalenia, które dokumenty z całego środowiska hipertekstowego (całego systemu WWW) wskazują na dane dwie strony, należące do kolekcji. Należy więc dotrzeć do stron (dokumenty d_k), które w swojej treści zawierają odsyłacze do stron, dla których obliczane jest podobieństwo (adresy URL strony d_i oraz d_j). W tym celu należy zadać wyszukiwarkom pytanie q_1 składające się iloczynu logicznego adresów URL strony d_i ($adresURL_i$) oraz d_j ($adresURL_j$), traktowanych jako słowa kluczowe[1]:

$$q_1 = adresURL_i \wedge adresURL_j.$$

Już na tym etapie pojawia się trudność związana z różną postacią adresów URL. Na stronie mogą występować adresy względne, które dopiero w połączeniu z adresem serwera dają pełny adres URL. W adresie może występować (albo nie) nazwa pliku na serwerze WWW. W tym drugim przypadku serwer przyjmuje pewne wartości domyślne[2]. Istnieją także alternatywne nazwy zarejestrowane w serwerach nazw domen (DNS) – różne dla tego samego serwera WWW[3]. W konsekwencji należałoby zadać nie pytanie q_1 a raczej zbiór pytań zawierający wszystkie możliwe kombinacje postaci adresu $adresURL_i$ z wszystkimi możliwymi postaciami adresu $adresURL_j$.

Po uzyskaniu odpowiedzi z wyszukiwarki należy pobrać z sieci Internet treść stron zwróconych jako odpowiedź wyszukiwarek, aby potwierdzić, czy rzeczywiście posiadają one odsyłacze do dwóch interesujących nas stron.

Jeżeli niezbędnym jest uzyskanie macierzy podobieństw, to wtedy koniecznym, jest zadanie po jednym pytaniu skierowanym do wyszukiwarek, dla każdej pary dokumentów z kolekcji (problem różnych postaci adresów URL został tutaj pominięty).

Daje to łącznie $\binom{N}{2} = \frac{N(N-1)}{2}$ pytań do wyszukiwarek, gdzie: N to liczba dokumentów w kolekcji D. Zachodzi tutaj także konieczność uzyskania treści wielu stron, zwróconych jako odpowiedzi na te pytania. Można to optymalizować, tj. zadawać pytanie o każdą stronę ($q_2 = adresURL_i$) czyli zadać tylko N pytań i wśród treści stron, których adresy są podane w odpowiedzi, szukać tych, które zawierają adres $adresURL_j$. Nie jest jednak pewne, czy będzie to rozwiązanie szybsze niż zadanie pierwszego pytania q_1, gdyż prawdopodobnie konieczne będzie pobranie treści większej liczby stron.

W związku z powyższym można zredukować wzór (3) poprzez usunięcie z niego liczb tych odsyłaczy, które są związane ze wspólnymi przodkami – współcytowaniem. W wyniku tego, powstanie nowa wersja D funkcji pierwotnego podobieństwa strukturalnego:

$$S_D^{ps}(d_i,d_j) = \begin{cases} \dfrac{nl_{ij} + nl_{ji} + \mu\left(nl_{ij}^{\Rightarrow} + nl_{ji}^{\Rightarrow}\right)}{nl_i^{\rightarrow} + nl_j^{\rightarrow}}, & \text{dla } nl_i^{\rightarrow} + nl_j^{\rightarrow} > 0 \\ 0, & \text{dla } nl_i^{\rightarrow} + nl_j^{\rightarrow} = 0 \end{cases} . \qquad (5)$$

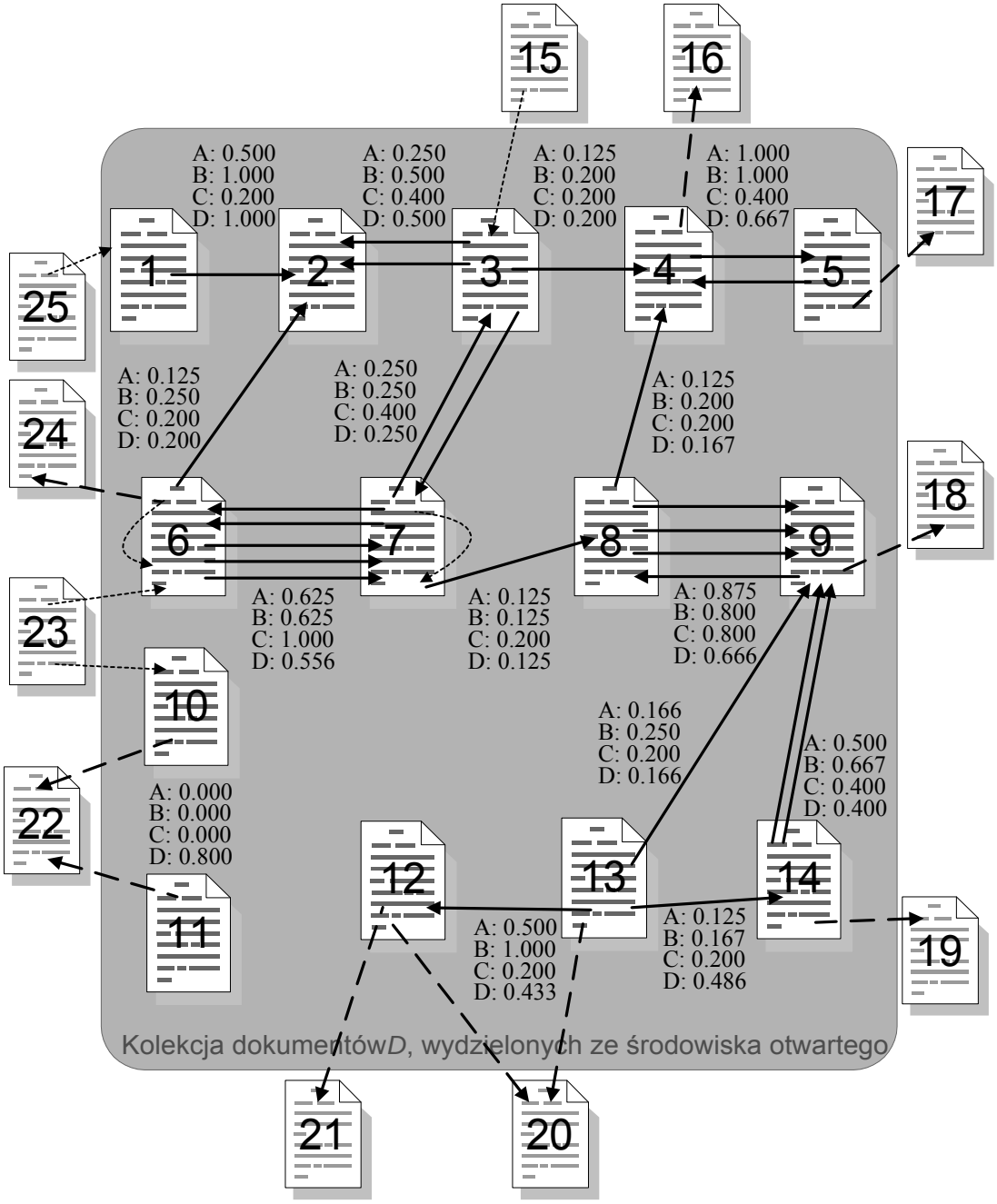
Podobnie jak w przypadku pozostałych wersji, zbiór wartości funkcji $S_D^{ps}(d_i,d_j)$ to przedział [0,1].

Po uwzględnieniu (2) oraz (4), otrzymujemy:

$$S_D^{ps}(d_i,d_j) = \begin{cases} \dfrac{\left(nl_{ij}^{s} + \lambda\, nl_{ij}^{n}\right) + \left(nl_{ji}^{s} + \lambda\, nl_{ji}^{n}\right) + \mu\left(nl_{ij}^{s\Rightarrow} + \lambda\, nl_{ij}^{n\Rightarrow} + nl_{ji}^{s\Rightarrow} + \lambda\, nl_{ji}^{n\Rightarrow}\right)}{nl_i^{\rightarrow} + nl_j^{\rightarrow}}, & \text{dla } nl_i^{\rightarrow} + nl_j^{\rightarrow} > 0 \\ 0, & \text{dla } nl_i^{\rightarrow} + nl_j^{\rightarrow} = 0 \end{cases}$$

5. WŁASNOŚCI PIERWOTNEGO PODOBIEŃSTWA STRUKTURALNEGO

Na rys. 1 przedstawiono kolekcję dokumentów hipertekstowych, dla których obliczono parami wartości funkcji podobieństwa dla różnych wersji tej funkcji (A, B, C oraz D). Założono, że wszystkie odsyłacze mają charakter semantyczny, czyli $nl_{ij}^{n} = nl_{ji}^{n} = 0$, co oznacza, że $nl_{ij} = nl_{ij}^{s}$, zaś $nl_{ji} = nl_{ji}^{s}$. Przy ustalaniu wartości podobieństwa dla wersji C konieczne było określenie wartości nl_{max}, która dla tej kolekcji wynosi 5 (liczba odsyłaczy między dokumentami 6 i 7). Dla wersji D, obliczeń dokonano wg wzoru (5), przy współczynniku μ=0.8.



Rys. 1. Kolekcja *D* dokumentów hipertekstowych z obliczonymi wartościami funkcji pierwotnego podobieństwa strukturalnego.

Odsyłacze bezpośrednie (łączące dwa dokumenty, należące do kolekcji *D*) narysowane są linią ciągłą. Odsyłacze oznaczone linią przerywaną to te, które biorą udział jedynie w obliczaniu podobieństwa $S_D^{ps}(d_i,d_j)$ i prowadzą do dokumentów nie należących do kolekcji. Są to odsyłacze: (4,16), (5,17), (6,24), (9,18), (10,22), (11,22), (12,21), (12,20), (13,20), (14,19).

Odsyłacze, które w ogóle nie były uwzględniane przy obliczaniu wartości funkcji pierwotnego podobieństwa strukturalnego (dla żadnej wersji) zostały oznaczone linią kropkowaną. Dotyczy to odsyłaczy prowadzących do innych części tego samego dokumentu (z d_i do d_i) oraz prowadzące z dokumentów nie należących do kolekcji D do dokumentów do kolekcji należących. Są to odsyłacze: (6,6), (7,7), (15,3), (25,1), (23,6), (23,10).

Dla funkcji podobieństwa obliczanej wg wersji D, w kolekcji z rys. 1, istnieje $S_D^{ps}(d_i,d_j)>0$, dla niektórych takich par (d_i,d_j), przy których podobieństwo liczone według pozostałych wersji jest równe 0. Jest to związane z tym, że nie ma odsyłaczy bezpośrednio łączących dokumenty w takiej parze ($nl_{ij}=nl_{ji}=0$). Między dokumentem 10 oraz 11 wartość pierwotnego podobieństwa strukturalnego $S_D^{ps}(10,11)=0.8$, zaś $S_A^{ps}(10,11)= S_B^{ps}(10,11)= S_C^{ps}(10,11)=0$. Istnieje więcej podobnych, niezerowych pierwotnych podobieństw strukturalnych liczonych wg wersji D, niezaznaczonych na rysunku: $S_D^{ps}(1,3)=0.267$, $S_D^{ps}(1,6)=0.267$; $S_D^{ps}(3,5)=0.267$, $S_D^{ps}(3,6)=0.533$; $S_D^{ps}(3,8)=0.2$; $S_D^{ps}(5,8)=0.267$; $S_D^{ps}(7,9)=0.32$; $S_D^{ps}(8,13)=0.4$; $S_D^{ps}(8,14)=0.571$. Dla dokumentów 3 oraz 8 pośrednie odsyłacze cytowania prowadzą nawet do dwóch, różnych, innych dokumentów 4 oraz 7.

Jeżeli dla dwóch dokumentów d_i i d_j nie istnieją bezpośrednie[4] ($nl_{ij}= nl_{ji}=0$) ani pośrednie[5] ($nl_{ij}^{\Rightarrow} = nl_{ji}^{\Rightarrow} = nl_{ij}^{\Leftarrow} = nl_{ji}^{\Leftarrow} =0$) odsyłacze je łączące, wtedy $S_A^{ps}(d_i,d_j)= S_B^{ps}(d_i,d_j)= S_C^{ps}(d_i,d_j)=0$.

Jeżeli obliczane jest podobieństwo danego dokumentu d_i do niego samego, wtedy otrzymujemy: $(\forall d_i \in D)\ S^{ps}(d_i,d_i) = 0$. Wynika to z tego, że odsyłacze odnoszące się do innych części tego samego dokumentu są pomijane w procesie obliczania pierwotnego podobieństwa strukturalnego, dla wszystkich jego wersji. W związku z tym liczniki ułamków mają wartość 0.

Pierwotne podobieństwo strukturalne spełnia warunek symetrii tj. $S^{ps}(d_i,d_j)=S^{ps}(d_j,d_i)$, dla wszystkich wersji, oraz (także dla wszystkich wersji) nie spełnia warunku trójkąta (opis odpowiednich warunków jest zawarty w [4, 11]).

W przeciwieństwie do innych propozycji podobieństwa strukturalnego zawartych w literaturze (zobacz przeglądy w [10, 11]), pierwotne podobieństwo strukturalne uwzględnia gęstość odsyłaczy w dokumentach, dla których dokonuje się obliczeń.

Pierwotne podobieństwo strukturalne spełnia więc założenia zawarte w pkt. 3. Wersje A, B, C spełniają punkty a), b), d), zaś wersja D dodatkowo punkt c).

Zbiorem wartości pierwotnego podobieństwa strukturalnego jest przedział [0,1].

6. ZASTOSOWANIE FUNKCJI PODOBIEŃSTWA STRUKTURALNEGO

Podobieństwo dokumentów jest szeroko stosowane w wyszukiwaniu informacji. Jest ono szczególnie istotne w automatycznym przetwarzaniu ogromnego, otwartego i szybko rozwijającego się środowiska WWW. Borodin i inni [2] wykorzystali je w analizie odsyłaczy (*hypertext link analysis*), którą zastosowali następnie do wyszukiwania informacji.

W pracy [11] opisaną tutaj funkcję pierwotnego podobieństwa strukturalnego, w wersji D ze wzoru (5), w połączeniu z podobieństwem tekstowym, zastosowano do grupowania dokumentów, w szczególności do tworzenia hierarchii grup - stron WWW, będących wynikiem wyszukiwania w wyszukiwarkach internetowych.

Podobieństwo strukturalne ma także istotne znaczenie przy automatycznej klasyfikacji, zwłaszcza przy tworzeniu katalogów a także w rankingach, czego przykładem jest jedna z najlepszych wyszukiwarek - Google.

7. PODSUMOWANIE

Odsyłacze są w dokumentach hipertekstowych naturalnym źródłem podobieństwa między dokumentami. O ich znaczeniu świadczy duża liczba artykułów naukowych związanych z odsyłaczami w środowisku WWW a także ich komercyjne wykorzystanie. Prowadzone są także prace związane ze zdefiniowaniem nowych standardów dla odsyłaczy w systemach hipermedialnych, czego przykładem jest XLink - język odesłań dla dokumentów XML, opublikowany czerwcu 2001 [12].

Oczywiście do wyznaczania podobieństwa - oprócz odsyłaczy - można także wykorzystać inne nośniki informacji, jak: treść tekstową, popularność, umiejscowienie lub w przypadku dokumentów hipermedialnych - elementy multimedialne [10]. Najlepsze rezultaty daje jednak połączenie wielu źródeł informacji.

8. LITERATURA

[1] Baron L, Tague-Sutcliffe J., Kinnucan M.T., Carey T.: *Labeled, typed links as cues when reading hypertext documents*. Journal of the American Society for Information Science, Volume 47, Number 12, December 1996, s. 896-908.

[2] Borodin A., Roberts G.O., Rosenthal J.S., Tsaparas P.: *Finding Authorities and Hubs From Link Structures on the World Wide Web*. The Tenth International World Wide Web Conference Proceedings, 2001, http://www10.org/cdrom/papers/314/index.html

[3] Chen C.: *Structuring and visualising the WWW by generalised similarity analysis*. W: Bernstein M., Carr L., Østerbye K. (eds.): Hypertext 97. The Eigth ACM Conference on Hypertext, University of Southampton, UK, ACM Press, 1997, s. 177-186.

[4] Dąbrowski M., Laus-Mączyńska K.: *Metody wyszukiwania i klasyfikacji informacji*. Wydawnictwa Naukowo-Techniczne; Warszawa, 1978.

[5] Frei H.P., Stieger D.: *The use of semantic links in hypertext information retrieval*. Information Processing & Management, Vol. 31, No. 1, 1995, s. 1-13.

[6] Garzotto F., Paolini P., Schwabe D.: *HDM - A Model-Based Approach to Hypertext Application Design*. ACM Transactions on Information Systems, Vol. 11, No. 1, January 1993, s. 1-26.

[7] Haas S.W., Grams E.S.: *Page and link classifications: connecting diverse resources*. W: Proceedings of the Third ACM Conference on Digital Libraries, June 23-26, 1998, Pittsburgh, PA, USA, ACM Press, 1998, s. 99-107.

[8] Kazienko P.: *Struktura hipertekstu a struktura systemu WWW*. Zagadnienia Informacji Naukowej, Nr 2 (72), 1998, s. 36-56. ftp://ftp.zsi.pwr.wroc.pl/publications/Kazienko/ZIN1998.

[9] Kazienko P.: *Rodzaje stron i odsyłaczy w systemie WWW*. Informatyka, Nr 2, Luty 1999, s. 24-35. ftp://ftp.zsi.pwr.wroc.pl/publications/Kazienko/Informatyka2-99.

[10] Kazienko P.: *Źródła podobieństwa stron WWW*. W: MiSSI 2000. Multimedialne i Sieciowe Systemy Informacyjne. Materiały konferencyjne pod red. Cz. Daniłowicza. Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław, 2000, s. 91-102. http://www.zsi.pwr.wroc.pl/missi2000/referat7.htm.

[11] Kazienko P.: *Grupowanie dokumentów hipertekstowych na podstawie drzewa maksymalnych przepływów. Praca doktorska*. Raport Zakładu Systemów Informacyjnych Politechniki Wrocławskiej PRE 31, 2000.

[12] Kazienko P.: *XLink - the Future of Document Linking*. Proceedings of the 23rd International Conference ISAT'2001, September 2001.

[13] Pitkow J.,Pirolli P.: *Life, death, and lawfulness on the electronic frontier*. W: Steven P. (ed.): CHI 97: Conference Proceedings on Human Factors in Computing Systems, Atlanta, Georgia, 22-27 March 1997. ACM/Addison-Wesley, 1997, s. 383-390.

Autor: Przemysław Kazienko

Zakład Systemów Informacyjnych, Wydział Informatyki i Zarządzania

Politechnika Wrocławska

emal: kazienko@pwr.wroc.pl

WWW: http://www.zsi.pwr.wroc.pl/pracownicy/kazienko/index.html

---

[1] Odpowiedzi wyszukiwarek będą niestety przybliżone, gdyż nie obejmują one całego środowiska hipertekstowego. Wg różnych badań pojedyncza wyszukiwarka nie pokrywa więcej niż 40-50% całego systemu WWW [11].

[2] zwykle jest to index.html lub default.html, ale nie zawsze

[3] Na przykład, wszystkie adresy: http://www.gsmsend.com/index.html, http://www.gsmsend.com/, http://gsmsend.com, http://www.gs.com, http://www.gsmsend.com/index.htm, http://www.oldgsmsend.com/index.html wskazują na ten sam dokument w sieci WWW.

[4] tj. dla wersji A, B oraz C

[5] dla wersji D