LOGICAL DOCUMENTS ON WWW

Martin MAYER, Master Degree Programme (5)
Dept. of Information Systems, FIT, BUT
E-mail: xmayer01@stud.fit.vutbr.cz

Supervised by: Ing. Radek Burget

ABSTRACT

Proposing a method of identifying logical documents in WWW data. Pages in WWW are sometimes designed for presentation and do not always reflect logical structure, while a logical document is a data unit representing logical structure. One logical document often corresponds to a connected sub-graph consisting of multiple pages. Therefore, for various WWW data processing that should capture logical structure, such as querying facilities, extended support for user navigation, and WWW structure analysis, logical documents are more appropriate data units than pages. This paper is about identifying just these logical documents on WWW.

1 ÚVOD

Významný rozdíl mezi WWW a dalšími druhy dat které byly navrhovány v tradičních databázových systémech je, že WWW je vytvořen mnoha uživateli, zatímco data v tradičních databázových systémech jsou vytvořena relativně malým množstvím lidí, nebo je přinejmenším uspořádal nějaký správce z dat vytvořených mnoho uživateli.

"Je vytvořen mnoha uživateli" v tomto kontextu má dva aspekty. Zaprvé, autoři zahrnují kdejakého uživatele, od profesionálů k nováčkům. Nováček může snadno tvořit WWW stránky a nynější WWW datový model poskytuje prostý mechanismus a dokonce s tak jednoduchým modelem, že nějaké mechanismy jsou často použité chybně. Zadruhé, uživatelé po celém světě tvoří stránky nezávisle bez nějaké centrální kontroly. Ačkoli je na WWW mnoho druhů sémantické struktury, následkem jsou modely dat vyjádřené jen implicitně a ve velmi různorodých formách.

Mezi různými druhy sémantické struktury, které jsou implicitně vyjádřené na WWW, se budeme soustředit na strukturu logických dokumentů. Logický dokument je spojený podgraf skládající se z rozmanitých stránek, které celkově odpovídají jednomu nezávislému kompletnímu dokumentu. WWW je obrovské množství hypertextových dat skládající se z obrovského množství WWW stránek a jediného druhu spojení, který je spojuje. Spojení(linky) mohou mít různé významy. Některé linky jsou užívány pro poskytnutí cesty skoku na další (nebo stejné) WWW stránky obsahující související nebo doporučené informace, a některé linky jsou umístěny autorem stránky jako standardní navigace procházející soubor stránek, které celkově tvoří jeden kompletní dokument. Často jsou

používány linky nabízející cestu navigace k prioritním stránkám, jako jsou odkazy na předchozí “back” stránku nebo odkazy na vrchol “top” stránky. Touto cestou, autoři stránek často dělí jeden dokument do souboru stránek a spojují je přes linky, u kterých se předpokládá, že budou tvořit standardní navigační cesty. Takový spojený podgraf skládající se z takového souboru stránek nazýváme tedy "logický dokument".

Zjišťování struktury logických dokumentů je velmi užitečné pro podporu prohledávání, dotazování, a porozumění WWW. Je to také velmi užitečné jako preprocessor pro různé další druhy analýz WWW dat.

Nový bod naší metody je, že prvně rozlišujeme cesty které jsou navrženy autorem stránky tak, aby byly standardními navigačními cestami a používáme výsledek pro objevování logických dokumentů. Po zjištění takových cest, můžeme určit logické dokumenty přesněji. V naší metodě užíváme tři druhy informací za účelem odhalit logické dokumenty: link struktura, adresářová struktura v URI, a četnost termů na každé stránce. První dva druhy informací jsou užívány pro identifikaci standardních navigačních cest, a třetí je užívaný pro rozdělávání grafu do podgrafů.

2 NALEZENÍ LOGICKÝCH DOKUMENTŮ

Odhalování logický dokumentů na WWW provádíme ve dvou krocích:

2.1 URČUJÍCÍ CESTY

Odhalování hierarchické struktury na stránkách zjišťováním určujících cest zamýšlených autory ve formě standardních navigačních cest z míst nejvrchnějších stránek k ostatním stránkám.

Z globálního pohledu, celé WWW stránky také mají obvykle hierarchickou strukturu. Většina WWW stránek se skládá z následujících elementů:

- vrchní stránka(y) pro přímý přístup
- ostatní stránky
- hierarchické cesty z vrchních stránek směřující k ostatním stránkám
- jiné druhy linků uvnitř stránek(sites)
- linky ke stránkám ostatních stránek(sites)

Jestliže extrahujeme jen linky pracující jako standardní cesty od vrcholu stránek ke každé stránce, struktura stránek je redukována na hierarchie. Na nějakých stránkách jsou stránky organizované do jedné hierarchie s jednou vrchní stránkou, a na dalších stránkách jsou více než dvě vrchní stránky a stránky jsou organizované do rozmanitých hierarchií mající kořeny právě vrchní zmíněné stránky. Jak bylo řečeno, každý logický dokument je také hierarchie skládající se z linků jako cest pro čtenáře k projití celého dokumentu.

Zjištění hierarchie na stránce se rovná rozpoznání cest z jejich vrchních stránek ke každé stránce. Tyto cesty nazýváme „určující cesty“ (route paths), protože jsou připraveny autory stránek nebo manažery WWW stránky jako standardní cesty pro všechny návštěvníky dané stránky. Přesněji definujeme určující cestu takto:

Definice: Určující cesta stránky *n* je cesta končící v *n* skrze kterou autor stránky *n* předpokládal, že jakýkoliv čtenář dosáhne *n* právě touto cestou.

Ačkoli definice určující cesty může být napsána takto, je tato definice ryze intuitivní. Není vždy jedinečná odpověď na otázku, zda-li cesta je určující cesta dané stránky a ani když se ptáme samotného autora, nemusí to být vždy jasné. Proto je těžké vyvinout metodu charakterizovat určující cesty dokonale. K určení jedné určující cesty k dané stránce je v naší práci použito několika heuristik, které analyzují hierarchii stránek a snaží se redukovat všechy příchozí linky na jeden jediný.

2.2 ROZDĚLENÍ NA ZÁKLÁDĚ PODOBNOSTI

Rozdělením výsledné hierarchie do podhierarchií odpovídající logickým dokumentům založených na podobnosti obsahu sousedních stránek.

Pro zjištění podobnosti stránek je použita technika TFxIDF. TFxIDF (term frequency, inverse document frequency) je jedna z technik získávání informací z dokumentů založená na vektorovém modelu prostoru, kde prostor slouží pro reprezentaci všech dokumentů a vektor pro určení jednoho dokumentu v prostoru. Tato technika je používána ke spočítání důležitosti termu(slova) objevujícího se v textu. Použití TFxIDF metody ve spojení s funkcí kosinové podobnosti Sim se zdá být vhodné pro použití v naší práci z hlediska dostačující přesnosti určení informací. Funkce kosinové podobnosti $Sim(d_i,d_j)$ měří kosinus úhlu mezi dvěma dokumenty reprezentovanými vektory d_i a d_j a vrací číslo mezi 0 (nesouvisející) a 1 (identické).

Postup získávání logických dokumentů: vypočítáme podobnost všech sousedních párů uzlů a opakovaně spojujeme uzly s větší podobností. Když proběhne algoritmus volaný na kořenový uzel stránky, dostaneme logické dokumenty. Jestli považujeme optimální rozložení grafu s váhami na hranách, pak jednodušší algoritmy, jako algoritmus, který opakovaně spojuje nejvíce si podobné sousedící uzly dokud počet uzlů není zredukován na nějaké číslo, vrací odpovědi, které jsou optimální jen v nějakém smyslu. V našich experimentech, náš algoritmus začíná od listů a proto dává priority listům a vrací lepší odpovědi.

3 ZÁVĚR:

V této práci jsme vyvinuli novou metodu objevování logických dokumentů na WWW. V této metodě objevujeme logické dokumenty ve dvou krocích: nejprve identifikujeme hierarchii stránek a pak ji rozdělíme na podhierarchie korespondující logickým dokumentům. Identifikováním hierarchie stránek můžeme objevovat logické dokumenty přesněji.

LITERATURA

[1] Tajima, K.: New Techniques for the Discovery of Logical Documents in WWW. In International Symposium on Database Applications in Non-Traditional Environments. 1999

[2] Mizuuchi, Y., Tajima, K., Tanaka, K.: Retrieval of graph structured data based on cut partitioning. IPSJ SIG Notes, 1997, 97-DBS-113:281–286

[3] Gibson, D., Kleinberg, J., Raghavan, P.: Inferring WWW communities from link topology. In Proc. of ACM Hypertext, pages 225–234, 1998

[4] Chakrabarti, S., Dom, B., Indyk, P.: Enhanced hypertext categorization using hyperlinks. In Proc. of ACM SIGMOD, pages 307–318, 1998