Metrické indexování vektorových modelů v oblasti information retrieval

Tomáš Skopal

Katedra informatiky, FEI, VŠB - Technická Univerzita Ostrava,
17. listopadu 15, 708 33, Ostrava-Poruba
Tomas.Skopal@vsb.cz

Abstrakt. Pro vektorové modely v oblasti Information Retrieval dosud nebylo navrženo efektivní indexovací schéma. Absence efektivní implementace vektorového modelu znemožňuje jeho použití pro velké kolekce dokumentů. V tomto příspěvku je shrnut současný stav problematiky a dále je zde představeno metrické indexování vektorových modelů, kterým se v rámci svého doktorského studia převážně zabývám.

1 Úvod

Oblast *Information Retrieval* [BaNe99] (dále IR, v českém prostředí oblast dokumentografických informačních systémů [PSH98]) se zabývá metodami vyhledávání v rozsáhlých kolekcích dokumentů. V minulosti byl kladen hlavní důraz na vyhledávání v kolekcích textů (tzv. Text Retrieval [Be99]). V současné době sílící multimedialní trend rozšiřuje hranice IR také do oblasti vyhledávání v multimediálních databázích, tj. vyhledávání v kolekcích dokumentů obsahujících XML, obrázky, audio, video, časové sekvence, DNA, atd.

Kvalitativním rámcem oblasti IR je návrh modelů pro vyhledávání. V současnosti existují tři klasické modely a jejich modifikace. Kvalitativně nejlepší výsledky (vedle boolovského a pravděpodobnostního modelu) podává vektorový model a jeho modifikace.

Na rozdíl od boolovského modelu, jehož implementace pomocí invertovaného seznamu je efektivní, a pravděpodobnostního modelu, který je víceméně teoretický, pro vektorové modely dosud nebylo navrženo efektivní indexovací schéma. Absence efektivní implementace vektorového modelu znemožňuje jeho použití pro velké kolekce dokumentů. S téměř exponenciálním nárůstem dokumentů dostupných např. na internetu je potřeba efektivního vyhledávání více než zřejmá.

Cílem mého snažení v rámci doktorského studia je návrh efektivní implementace vektorových modelů pomocí metrického indexování pro velké kolekce dokumentů.

2 Vektorový model

V klasickém Text Retrieval vektorovém modelu je každý dokument kolekce reprezentován vektorem vah termů (klíčových slov) v dokumentu. Hodnota každé souřadnice (tj. váha příslušného termu) specifikuje nakolik je daný term relevantní

vůči obsahu dokumentu. Váhy mohou být konstruovány nejrůznějším způsobem, nejpopulárnější je konstrukce pomocí *inverzní frekvence termu v dokumentech* [Be99]. Soubor vektorů dokumentů tvoří matici termů-dokumentů (každý sloupec matice reprezentuje jeden dokument, viz obr. 1), která reprezentuje původní kolekci dokumentů a na které také probíhá vyhledávání, resp. vyhodnocování uživatelských dotazů.

dokument / term \	D_1	D_2	D_3	D_4	D_5
database	0	0.48	0.05	0	0.70
vector	0.23	0	0.23	0	0
index	0.43	0	0	0	0
image	0	0	0.10	0	0.54

Obr. 1. Matice termů-dokumentů obsahující pět vektorů dokumentů (sloupce matice).

Poslední, a snad i nejdůležitější věcí, kterou je potřeba speficikovat pro vektorový model, je *funkce podobnosti*, která dvěma vektorům přiřadí míru podobnosti. Tato funkce musí být konstruována tak, aby reflektovala skutečnou podobnost dokumentů vnímanou lidským subjektem. Taková funkce podobnosti umožňuje jak porovnávat dokumenty vůči sobě, tak vyhodnotit uživatelský dotaz, který lze reprezentovat vektorem stejně jako dokument. Dnes nejpopulárnější funkcí podobnosti je *kosinová míra*.

2.1 Indexování latentní sémantiky

Velmi zjednodušeně řečeno, model indexování latentní sémantiky (LSI) je algebraickým rozšířením vektorového modelu o fázi, kdy je matice termů-dokumentů pomocí singulárního rozkladu (SVD) redukována na matici konceptů-dokumentů. Vektor konceptu (singulární vektor) je tvořen váhami původních termů. Matice konceptů-dokumentů je pak vyjádřena v bázi konceptů. Vektor dokumentu (zde se nazývá vektorem pseudo-dokumentu) je vyjádřen jako lineární kombinace všech uvažovaných konceptů. Díky bázi konceptů, která je vytvořena statisticky z původní matice termů-dokumentů, má LSI několik výhod:

- Odhaluje latentní (skrytou) sémantiku dokumentů. Nalezené koncepty (jejich vektory) lze totiž uspořádat podle důležitosti. Koncept si můžeme představit jako určité téma, které se vyskytuje společně v několika (mnoha) dokumentech. Zanedbávají se ty koncepty, které se vyskytují v malém počtu dokumentů a tedy z hlediska celé kolekce působí jako šum.
- Zanedbáním nesignifikantních konceptů lze škálovatelně redukovat dimenzi prostoru dokumentů a tím i dimenzi výsledné matice konceptů-dokumentů.
- Z teorie se dá předpokládat (a také jsme to experimentálně ověřili [5]), že vektory pseudo-dokumentů budou lépe shlukovány (vzhledem ke kosinové

míře), než vektory dokumentů klasického vektorového modelu. Tato skutečnost je výhodná pro hierarchické metrické indexování, kterým se zabýváme (viz kapitola 3).

Nevýhodou současného LSI je (vedle výpočetní náročnosti konstrukce matice konceptů-dokumentů) také fakt, že se nejedná o indexování v pravém slova smyslu. Konstrukcí matice konceptů-dokumentů a projekcí dotazovacího vektoru úloha LSI končí a dotazy se vyhodnocují nad maticí konceptů-dokumentů naprosto stejně jako nad maticí termů-dokumentů v případě klasického vektorového modelu.

2.2 Vyhodnocování dotazů

Vyhodnocování dotazů je hlavním praktickým problémem implementace vektorového modelu. Při naivní implementaci sekvenčním průchodem matice je dotaz (i ten nejjednodušší) vyhodnocen tím, že je vektor dotazu porovnán pomocí funkce podobnosti se všemi vektory dokumentů, tj. celá matice musí být postupně načtena a musí být provedeno tolik výpočtů funkce podobnosti, kolik je dokumentů v kolekci (resp. sloupců v matici). Uvědomíme-li si, že velikost matice pro velké kolekce může být reálně i v desítkách gigabajtů, sekvenční průchod je prakticky nepoužitelný. Také milión volání funkce podobnosti není levná záležitost, bereme-li v úvahu velikost dimenze srovnávaných vektorů.

Jako příklad si vezměme kolekci miliónu dokumentů obsahující cca 100.000 unikátních termů. Velikost příslušné matice bude 10^{11} reálných hodnot. V praxi je tato matice řídká (max. 1% hodnot je nenulových), neboť každý dokument obsahuje jen malou podmnožinu termů. Přesto je k uložení této matice (ve formátu CCS [Be99]) potřeba cca 7 GB. V případě LSI je sice dimenze redukována na několik set (např. na 300), ovšem matice konceptů-dokumentů již není řídká a její uložení si v tomto případě vyžádá téměř 300 MB.

Pro implementaci vektorového modelu byly v minulosti použity signaturové metody [MPS03] (signaturové soubory, váhově rozdělené signaturové soubory, S-stromy, VA-files, atd.), ovšem z hlediska efektivního indexování textových dokumentů se jejich použití ukázalo jako nedostatečné.

Je zřejmé, že pro efektivní implementaci vektorového modelu je potřeba nad danou maticí vytvořit index, pomocí něhož by se při dotazu přistupovalo pouze k malé části matice.

3 Metrické indexování

V našem přístupu jsme chtěli vytvořit nad maticí index využitím datových struktur z oblasti indexování prostorových dat [BBK01], z nichž některé jsou vhodné pro vyhodnocování dotazů vektorového modelu. K realizaci této ambiciózní představy je ovšem třeba počítat s typickými problémy v oblasti indexování vektorových dat. Všechny prostorové indexovací struktury (např. R-stromy, UB-stromy,

X-stromy, atd.) více či méně trpí tzv. prokletím dimenzionality, které se projevuje zejména neefektivním vykonáváním dotazů v případě indexování dat vysoké dimenze ("vysoké" je zde myšleno $d > 20$).

Pro indexování matice vektorového modelu jsme zvolili M-strom [CPZ97,Pa99] (námi revidovaný v [6]) – datovou strukturu umožňující indexovat obecné metrické prostory. Hlavními důvody pro použití M-stromu jsou jednak vyšší odolnost vůči prokletí dimenzionality a jednak jeho struktura, která je vytvářena přímo pro potřebu vyhledávání podobných dokumentů (similarity search), tj. pro realizaci přesně těch dotazů, které využívá vektorový model.

Předpokladem pro úspěšné použití M-stromu – jehož hierarchie dělí prostor na metrické regiony, ve kterých jsou poshlukovány podobné dokumenty – je reálná existence shluků v kolekci vektorů dokumentů. Předběžné výsledky jsme publikovali v [5] a [8], kde se tento předpoklad experimentálně potvrdil.

4 Další cíle

V další práci se chceme zaměřit jednak na další vylepšení M-stromu (jak jsme učinili již v [6]), jednak bychom chtěli vyvinout novou metrickou strukturu vycházející z M-stromu. Neméně důležitým cílem pro samotné indexování vektorových modelů bude volba použití vhodných metrik a dále pak použití podobnost zachovávajících semi-metrik, které se ukázalo v předběžných experimentech jako slibné pro velmi efektivní přibližné vyhledávání.

Granty a ocenění

Účastil jsem se řešení těchto grantů:
GAČR 201/00/1031, GAČR 201/03/1318, GAČR 201/03/0912.

Jako jeden z oceněných Ph.D. studentů v rámci "VLDB East Europe support program" jsem se účastil 29. ročníku prestižní mezinárodní konference Very Large Data Bases 2003 v Berlíně.

Ostatní aktivity

V rámci seminářů DIS (ČVUT, MFF UK Praha) a ARG (VŠB-TU Ostrava) jsem přednesl několik příspěvků z oblasti indexování multimediálních dat a information retrieval.

Byl jsem členem organizačních výborů DATESO 2002, CLA 2002, ZNALOSTI 2003 a DATESO 2003. Jsem členem Amphora Research Group (ARG) při katedře informatiky FEI.

Publikoval (nebo spolupublikoval) jsem 23 článků ve sbornících konferencí a v odborných časopisech, viz následující reference.

Reference

1. Skopal T., Krátký M., Snášel V., Pokorný J.: A New Range Query Algorithm for the Universal B-trees, submitted to *EDBT 2004*, Crete, Greece.
2. Barashev D., Krátký M., Skopal T.: Modern Approaches to Indexing XML Data, to appear in *Sborník vědeckých prací VŠB-Technická univerzita Ostrava*, VŠB-TU Ostrava, 2003.
3. Krátký M., Skopal T., Snášel V.: Multidimensional Term Indexing for Efficient Processing of Complex Queries, to appear in *Kybernetika Journal*, Institute of Information Theory and Automation of the Academy of Sciences of Czech Republic.
4. Krátký M., Skopal T., Snášel V.: Image Compression Using Space-Filling Curves, *ITAT 2003*, High Tatras, Slovakia
5. Skopal T., Moravec P., Krátký M., Snášel V., Pokorný J.: An Efficient Implementation of the Vector Model in Information Retrieval. In *5th National Russian Research Conference on Digital Libraries, RCDL*, St. Petersburg, Russia, 2003.
6. Skopal T., Pokorný J., Krátký M., Snášel V.: Revisiting M-tree Building Principles. In *ADBIS 2003*, LNCS, Springer-Verlag, Dresden, Germany, 2003.
7. Krátký M., Skopal T., Snášel V.: Efektivní vyhledávání v kolekcích obrázků tváří, *DATAKON 2003*, Brno.
8. Skopal T., Krátký M., Snášel V.: Efektivní implementace vektorového modelu pro dokumentografické informační systémy, *DATAKON 2003*, Brno.
9. Dvorský J., Krátký M., Skopal T., Snášel V.: Benchmarking the Multidimensional Approach for Term Indexing, *DATESO 2003*, Desná-Černá Říčka.
10. Krátký M., Skopal T.: Benchmarking the UB-tree, *DATESO 2003*, Desná-Černá Říčka.
11. Dvorský J., Krátký M., Skopal T., Snášel V.: Term Indexing in Information Retrieval Systems, *CIC'03*, Las Vegas, Nevada, USA, CSREA Press, 2003.
12. Skopal T., Snášel V., Krátký M., Svátek V.: Searching the Internet Using Topological Analysis of Web Pages, *CIC'03*, Las Vegas, Nevada, USA, CSREA Press, 2003.
13. Krátký M., Skopal T., Snášel V.: Vícerozměrný přístup pro netriviální vyhledávání termů, *ZNALOSTI 2003*, Ostrava.
14. Krátký M., Pokorný J., Skopal T., Snášel V.: Geometric framework for indexing and querying XML documents, *EurAsia-ICT 2002*, Springer-Verlag LNCS 2510, Shiraz, Iran.
15. Krátký M., Pokorný J., Skopal T., Snášel V.: The geometric approach for indexing XML data, *DATAKON 2002*, Brno.
16. Skopal T., Krátký T., Snášel V.: Geometrické indexování a dotazování multimediálních dat, *DATAKON 2002*, Brno.
17. Skopal T., Snášel V., Krátký M.: Image Recognition Using Finite Automata, *Prague Stringology Conference 2002*, Prague.
18. Skopal T., Krátký M., Snášel V.: Properties Of Space Filling Curves And Usage With UB-trees, *ITAT 2002*, Malino Brdo, Slovakia.
19. Snášel V., Skopal T., Ďuráková D.: Navigation Through Query Result Using Concept Order, *ADBIS 2002*, Research Communications, Bratislava, Slovakia.
20. Skopal T.: ACB Compression Method and Query Preprocessing in Text Retrieval Systems, *DATESO 2002*, Desná-Černá Říčka.
21. Krátký M., Skopal T., Snášel V.: Porovnání některých metod pro vyhledávání a indexování multimediálních dat, *Kybernetika - história, perspektívy, teória a prax*, Žilina, Slovakia, 2002.
22. Skopal T.: Architektura GIS z pohledu toků dat, *GIS OSTRAVA 2002*, Ostrava.
23. Skopal T.: Informační systém řízený toky dat, *OBJEKTY 2001*, Prague.

Reference

[BaNe99] Baeza-Yates R., Ribeiro-Neto B.: *Modern Information Retrieval*, Addison Wesley, New York, 1999.

[Be99] Berry M., Browne M.: *Understanding Search Engines, Mathematical Modeling and Text Retrieval*. Siam, 1999.

[BBK01] C. Böhm, S. Berchtold, and D. Keim. Searching in High-Dimensional Spaces -Index Structures for Improving the Performance of Multimedia Databases. *ACM-Computing Surveys*, 33(3):322-373, 2001.

[CPZ97] Ciaccia P., Pattela M., Zezula P.: M-tree: An Efficient Access Method for Similarity Search in Metric Spaces, *Proceedings of 23rd International Conference on VLDB*, Athens, Greece, 1997.

[MPS03] Moravec P., Pokorný J., Snášel V.: Vector Query with Signature Filtering, *Proceedings of the 6th Bussiness Information Systems Conference*, Colorado Springs, USA, 2003.

[Pa99] Patella M.: *Similarity Search in Multimedia Databases*, PhD thesis, Dipartmento di Elettronica Informatica e Sistemistica, Bologna, 1999.

[PSH98] Pokorný J., Snášel V., Húsek D.: *Dokumentografické informační systémy*. Karolinum, Praha, 1998.

Annotation

The vector models in the area of information retrieval still lack an efficient indexing schema. Owing to that fact, the vector models are not applicable for huge collections of documents. In this paper the current state of the art is summarized and the metric indexing of the vector models is proposed.