Információ-visszakereső módszerek egységes keretrendszere és alkalmazásai

Doktori (PhD) értekezés tézise

Kiezer Tamás

Témavezető: Dr. Dominich Sándor†
(1954 - 2008)

Pannon Egyetem

Műszaki Informatikai Kar

Informatikai Tudományok Doktori Iskola

2010

1 Tartalmi összefoglaló

Az Internet és a World Wide Web megjelenése mind gyakorlati, mind elméleti szempontból jelentős mértékben növelte az információ-visszakeresés fontosságát. Sokféle visszakereső módszer került kidolgozásra az elmúlt fél évszázad során, melyeket ma is folyamatosan fejlesztenek tovább.

A klasszikus módszerek egyike a vektortér módszer (Vector Space Model – VSM). Már két évtizede tudjuk, hogy a VSM nem vezethető le következetesen azon matematikai fogalmakból, melyeken alapszik, de ezidáig nem született megfelelő megoldás a problémára. Disszertációmban egy egységes, következetes, formális információ-visszakereső keretrendszert adok meg és bemutatom, hogy ennek alkalmazásával az általánosított vektortér módszer (Generalised Vector Space Model – GVSM), az LSI módszer (Latent Semantic Indexing model) és a VSM helyes matematikai formalizmust kap, amely konzisztens a gyakorlattal.

Az egységes keretrendszerben új, konzisztens visszakereső módszereket adok meg: az entrópia- és valószínűség-alapú módszert, valamint a kifejezetten Webes információ-visszakeresésre használható kombinált fontosság-alapú módszert. Utóbbit a WebCIR Webes keresőmotorban implementáltuk, mely szintén bemutatásra kerül a dolgozatban.

A megadott módszerek relevancia-hatékonyságát kísérleti úton vizsgáltam meg. Az entrópia- és valószínűség-alapú módszerek in vitro kiértékelése során 5 és 19 százalék közti javulás volt mérhető a VSM és LSI módszerekkel szemben. A WebCIR keresőmotor in vivo tesztelése során kapott eredmények alapján – a Yahoo!, Altavista, és MSN kereskedelmi keresőmotorok eredményeivel összehasonlítva – mondhatjuk, hogy a WebCIR visszakereső és rangsoroló technológiája versenyképes alternatívát jelent.

2 Tézisek

Az értekezés új tudományos eredményei az alábbiakban foglalhatók össze:

1. *Információ-visszakereső módszerek egységes keret-rendszere*

Az információ-visszakeresésre adott definíciókat megvizsgálva észrevehetjük, hogy azok nem különböző interpretációi az IR-nek, hanem nagyon hasonlóak. Ezt alapul véve megadtam az információ-visszakeresés egységes formális keretrendszerét.

(a) Megadtam a visszakeresés elvének matematikai mértékelméleten alapuló definícióját. A dokumentumokat (és a keresőkérdéseket) a fuzzy halmazelmélet segítségével határoztam meg [Chapter 3.2]. Majd a visszakeresést, mint két fuzzy halmaz metszetének számosságával definiált függvényt tekintettem [Lemma 4.1].

(b) Megmutattam, hogy az így megadott egységes keretrendszerben, az általánosított vektortér-modellt, a rejtett szemantikus indexelést (LSI) és a klasszikus vektortér-modellt újradefiniálva azok helyes

matematikai formalizmust kapnak, melyek konzisztensek a gyakorlattal [Chapter 4].

2. Entrópia- és valószínűség alapú visszakereső módszerek

Az új mértékelméleti megközelítés lehetővé teszi további, új és az elmélettel konzisztens visszakereső módszerek megadását. A fuzzy entrópiát és a fuzzy valószínűséget alapul véve új visszakereső módszereket adtam meg, melyek konzisztensek a matematikai hátterükkel.

(a) A visszakereső függvényben a fuzzy entrópiát véve mértéknek megadtam az Entrópia-alapú visszakereső módszert [Chapter 5.1].

(b) A visszakereső függvényben a fuzzy valószínűséget véve mértéknek megadtam a Valószínűség-alapú visszakereső módszert [Chapter 5.2].

(c) A módszerek relevancia-hatékonyságát sztenderd teszt-kollekciókon mértem. A gyakorlati eredmények alapján a VSM és LSI módszerekéhez képest átlagosan 5% és 19% közti hatékonyság növekedést tapasztaltam, mely azt mutatja, hogy a mértékelméleti megközelítésen alapuló egységes keretrendszer jó

alapja lehet új és hatékony visszakereső módszerek kifejlesztésének [Chapter 5.3].

3. Kombinált fontosság-alapú Webes információ-visszakereső módszer

A World Wide Web speciális tulajdonságai miatt a modern webes keresők jellemzően olyan visszakereső módszereket használnak, melyek részben klasszikus visszakereső módszereken, részben pedig a Webgráf speciális tulajdonságain alapulnak.

(a) Az (1.a) és (2.b) tézispontokban megfogalmazott keretrendszert és valószínűség alapú módszert használva új webes információ-visszakereső módszert adtam meg, mely tartalmi- és link alapú fontosságon, valamint hasonlóságon alapul. A módszert a WebCIR nevű keresőmotorban implementáltam [Chapters 6.4 and 7].

(b) A WebCIR kereső relevancia-hatékonyságának kiértékelésére 4 különböző módszert alkalmaztam, majd az eredményeket az Altavista, Yahoo!, és MSN keresők eredményeivel hasonlítottam össze. A kísérletek eredményei azt jelzik, hogy a Kombinált fontosság-alapú Webes visszakereső módszer versenyképes alternatívát jelenthet [Chapter 7.7].

3 Publikációk

Az értekezés témájához közvetlenül kapcsolódó publikációk

[P1] DOMINICH, S., **KIEZER, T.**, ERDÉLYI, M. (2008). WebCIR: Web ranking and search engine using combined method. *Studies on information and knowledge processes 13.* Infota, pp.: 53-74. **[thesis 2, 3]**

[P2] DOMINICH, S., **KIEZER, T.** (2007). A Measure Theoretic Approach to Information Retrieval. *Journal of the American Society for Information Science and Technology*. John Wiley & Sons, Vol. 58, no 8, pp.: 1108-1122, ISSN 1532-2882, IF=1.773. **[thesis 1, 2]**

Az értekezést megelőző, azt megalapozó publikációk

[P3] DOMINICH, S., GÓTH, J., **KIEZER, T.** (2006). Web-based Neuroradiological Information Retrieval System using three methods to satisfy different user's aspect. *Computerized Medical Imaging and Graphics*, ISSN 0895-6111, pp: 263-272, IF=1.090.

[P4] DOMINICH, S., **KIEZER, T.** (2005). Hatványtörvény, „kis világ” és magyar nyelv. *Alkalmazott Nyelvtudomány*, pp: 5-25, ISSN 1587-1061.

[P5] DOMINICH, S., GÓTH, J., **KIEZER, T.** (2005). NeuRadIR: A Web-Based NeuroRadiological Information Retrieval System. *ERCIM News*, vol. 61., pp:52-53, ISSN 0926-4981.

[P6] DOMINICH, S., GÓTH, J., M. HORVÁTH, **KIEZER, T.** (2005). 'Beauty' of the World Wide Web – Cause, Goal, or Principle. *Lecture Notes in Computer Science*, Springer Verlag, Volume 3408/2005, pp:67-80, ISSN 0302-9743, IF=0.515.

[P7] DOMINICH, S., GÓTH, J., **KIEZER, T.**, SZLÁVIK, Z. (2004). Entropy-based interpretation of Retrieval Status Value-based Retrieval, and its application to the computation of term and query discrimination value. *Journal of the American Society for Information Science and Technology*. John Wiley & Sons, Vol. 55, no 7, pp: 613-627, ISSN 1532-2882, IF=1.773.

Hivatkozások

[C1] Bujdosó, I. (2006) Rangado – vortstatistika ekzamenado de la plurlingva teksto de la konstitucipropono de Europa Unio. *Proceedings Internacia Kongresa Universitato Florenco*, Italio, 29 julio – 5 augusto, pp: 134-143

[C2] Ianeva, T., Boldareva, L., Westerweld, T., Cornacchia, R., Hiemstra, D., and de Vries, A.P. (2004). Probabilistic approaches to video retrieval. *Proceedings of TRECVID International Conference*, National Institute of Standards, NIST, USA, pp: 1-10

[C3] Lafouge, T., Prime-Claverie, C. (2005). Production and use of information. Characterization of informetric distri-butions using effort function and density function. Exponential informetric process. *Information Processing and Management*, vol. 41, pp: 1387-1394, Elsevier, ISSN 0306-4573, IF=1,295

[C4] Janssens, F., Leta, J., Glanzel, W., Moor, B. (2006). Towards mapping library and information science. *Information Processing and Management.* Elsevier, vol 42, no 2, pp: 1614-1642. ISSN 0306-4573, IF=1,215

[C5] Bordogna, G., Pagani, M., Pasi, G. (2006). A dynamic hierarchical fuzzy clustering algorithm for information filtering. *Studies in Fuzziness and Soft Computing*, Springer, vol. 197, pp: 3-23, ISSN 1434-9922.