Z Wikipedie, otevřené encyklopedie
ikona
Tento článek potřebuje úpravy.
Můžete Wikipedii pomoci tím, že ho vylepšíte. Jak by měly články vypadat, popisují stránky Vzhled a styl, Encyklopedický styl a Odkazy.

Konkrétní problémy: *celková wikifikace, formát dle Wikipedie:ES*

Získávání informací (anglicky *information retrieval*; IR) je činnost, během které člověk získává ze sbírky informačních zdrojů relevantní informace. Vyhledávání mohou být založena na fulltextových nebo na dalších obsahových parametrech. Získávání informací je vědecká disciplína zaobírající se hledáním informací v dokumentu, vyhledáváním samotných dokumentů a také vyhledáváním metadat, která popisují data, a databází textů, obrázků nebo zvuků.

Automatické systémy získávání informací pomáhají snižovat „informační zahlcení". Mnoho univerzit a veřejných knihoven využívá takové systémy k zajištění přístupu k různým knihám, časopisům a jiným dokumentům. Nejviditelnější službou získávání informací jsou webové vyhledávače.

Přehled

Proces získávání informací začíná momentem, kdy uživatel zadá dotaz do systému. Dotazy jsou formálními zápisy požadavku na informace, příkladem mohou být textové řetězce ve webových vyhledávačích. V procesu získávání informace nemusí dotaz jednoznačně identifikovat právě jeden objekt v souboru. Namísto toho může dotazu odpovídat vícero objektů více či méně relevantních.

Objekt je entita, která je reprezentovaná informací v souboru obsahů nebo databáze. Uživatelské dotazy jsou porovnávány s informacemi v databázi. Na rozdíl od klasických SQL dotazů z databází mohou i nemusí výsledky získávání informací přesně odpovídat dotazu, proto jsem většinou ohodnoceny. Toto hodnocení výsledků vytváří klíčový rozdíl mezi hledáním získávání informace a databázovým hledáním.[1]

Datovými objekty mohou být například textové dokumenty, obrázky, audio, myšlenkové mapy nebo videa. Často nejsou dokumenty uchovávány nebo ukládány přímo do systému získávání informací, ale jsou zde reprezentovány zástupci či metadaty.

Většina systémů získávání informací vypočítá ke každému objektu skóre podle toho, jak se shoduje s hledaným dotazem, a seřadí je podle těchto hodnot. Objekty s nejvyšším skóre jsou pak zobrazeny uživateli. Celý proces může být zopakován, pokud si to uživatel přeje.

Historie

Využívat počítače ke hledání relevantních informací bylo myšlenkou, kterou zpopularizoval článek *As We May Think* od Vannevar Bush v roce 1945. Zdá se, že byl Bush inspirován patenty na statistický stroj, se kterým přišel ve 20. a 30. letech Emanuel Goldberg. Statistický stroj hledal dokumenty uložené na filmovém pásu. První popis počítačového hledání informací představil Holmstrom v roce 1948 a informoval rovněž počítači Univac. Automatizovaný systém získávání informací byl představen v 50. letech: jeden se dokonce objevil v romantické komedii Desk Set z roku 1957. V 60. letech byla založena první velká výzkumná skupina zaměřující se na získávání informací, iniciátorem jejího vzniku byl Gerard Salton na Cornellově univerzitě. Do 70. let se objevilo několik různých technik získávání informací, které měly velmi slušné výsledky na menších korpusech textů, jakými byla například Cranfield collection (čítající několik tisíc dokumentů). Velké systémy (jako například Lockheed Dialog system) se začaly používat na začátku 70. let.[2]

V roce 1992 spolusponzoroval US Department of Defense společně s National Institute of Standards and Technology (NIST) setkání Text Retrieval Conference (TREC), které bylo součástí TIPSTER text program. Cílem tohoto setkání bylo nahlédnout komunitu vzniklou okolo problematiky získávání informací. To vyvolalo výzkum metod, které se zaměřovaly na obrovské korpusy. Příchod webových vyhledávačů ještě více posílil potřebu po rozsáhlých vyhledávacích systémech.

Typy modelů

Pro efektivní získávání relevantních dokumentů pomocí strategií získávání informací jsou dokumenty většinou transformovány do vhodné reprezentace. Každá „získávací” strategie obsahuje konkrétní model pro reprezentaci dokumentů.

První dimenze: matematický základ

- Množinové modely reprezentují dokumenty jako množiny slov či frází. Podobnosti jsou obvykle odvozeny z množinových operací těchto struktur. Běžné modely jsou:
  - Standardní booleovský model
  - Rozšířený booleovský model
  - Fuzzy retrieval
- Algebraické modely reprezentují dokumenty a dotazy jako vektory, matice nebo n–tice. Skalární hodnoty potom odpovídají podobnosti vektoru dotazu a vektoru dokumentu.
  - Vektorový model
  - Zobecněný vektorový model
  - (Rozšířený) tematický vektorový model
  - Rozšířený booleovský model
  - Latentní sématinická analýza
- Pravděpodobnostní modely přistupují k procesu výběru dokumentu jako k pravděpodobnostnímu odvozování. Podobnost je vypočítána z pravděpodobnosti, že dokument odpovídá danému dotazu. V těchto modelech se často využívají tvrzení z teorie pravděpodobnosti.
  - Binárně nezávislý model
  - Pravděpodobnostní modely relevantnosti na kterých je založena např. funkce BM25.
  - Odvozování nejistoty
  - Lingvistická analýza
  - Model divergence-from-randomness
  - Latentní Dirichletova alokace
- Modely typu feature-based retrieval

Druhá dimenze: vlastnosti modelu

Modely bez termínové provázanosti přistupují k různým termínům/slovům nezávisle. Ve vektorových modelech to znamená předpoklad ortogonality termínových vektorů, v pravděpodobnostních modelech předpoklad nezávislosti termínových proměnných. Modely s provázaností neoddělitelných termínů umožňují reprezentovat provázanost mezi termíny. Ovšem model sám definuje míru provázanosti dvou termínů. To je obvykle přímo či nepřímo odvozeno z toho, jak se objevují spolu v dokumentech v korpusu. Modely s transcendentní provázaností termínů také umožňují reprezentovat provázanost mezi termíny, ale nepředepisují provázanost mezi dvěma termíny. Míru provázanosti určují externí zdroje (např. lidé nebo specializované algoritmy).

Časová osa

- 19. století
  - 1801: Joseph Marie Jacquard vynalézá ruční tkalcovský stav s žakárovým ústrojím, což je první přístroj, který využívá děrné štítky. Díky nim se dají kontrolovat a ovládat jednotlivé nitě a jednoduše tak vytvářet složité vzory.
  - 80. léta 19. století: John Shaw Billings a Herman Hollerith vynalezli počítačový stroj pracující s děrnými štítky, který umožňoval rychle zpracovávat mnoho dat.
  - 1890: Shawovy a Hollerithovy děrné štítky a děrnoštítkový stroj jsou využity pro sčítání lidu ve Spojených státech
- 20. – 30. léta 20. století

Emanuel Goldberg předkládá patenty na "Statistical Machine” (tj. statistický stroj), což je vyhledávač dokumentů, který využívá fotoelektrické buňky a rozpoznávací paterny k vyhledávání na rolích mikrofilmových dokumentů.

- 40. – 50. léta
  - konec 40. let: Americká armáda čelí problému, jak indexovat The US military confronted problems of indexing and retrieval of wartime scientific research documents captured from Germans.
  - 1945: Článek As We May Think Vannevara Bushe vychází v Atlantic Monthly.

    - 1947: Hans Peter Luhn (výzkumný inženýr v IBM od roku 1941) začíná pracovat na mechanizovaném systému děrných štítků, který pomáhá vyhledávat chemické sloučeniny.
    - 50. léta: Rostoucí obavy o technologické zaostávání USA se SSSR podporují financování vědy, což napomáhá vyvíjet například mechanizované vyhledávací systémy literatury (Allen Kent a skupina okolo něj) či indexování citací (Eugene Garfield).
    - 1950: Termín "information retrieval” (tj. získávání informací) byl poprvé použit Calvin Mooers.
    - 1951: Philip Bagley provedl v magisterské práci na MIT první experiment v oblasti počítačového vyhledávání dokumentů.
    - 1955: Allen Kent nastoupil na Case Western Reserve University a stal se poté zástupcem ředitele Center for Documentation and Communications Research.
    - 1958: International Conference on Scientific Information Washington DC included consideration of IR systems as a solution to problems identified.
    - 1959: Hans Peter Luhn publikovat text Auto-encoding of documents for information retrieval.
- 60. léta
    - počátek 60. let: Gerard Salton začal na Harvardu pracovat v oblasti získávání informací, později pokračoval na Cornell
    - 1960: Melvin Earl Maron and John Lary Kuhns zveřejnili text "On relevance, probabilistic indexing, and information retrieval" v časopise Journal of the ACM; Cyril W. Cleverdon zveřejnil své první poznatky z Cranfield studies, které rozvíjejí model pro hodnocení systémů na získávání informací; Kent zveřejnil text Information Analysis and Retrieval.
    - 1963: Weinbergova zpráva "Science, Government and Information" vyjádřila myšlenku „krize vědecké informace.” Zpráva je pojmenována po Alvinovi Weinbergovi; Joseph Becker a Robert M. Hayes zveřejnili text týkající se získávání informací.
    - 1964: Karen Spärck Jones dokončila svou doktorskou práci na Cambridge s názvem Synonymy and Semantic Classification a pokračovala s prací v oblasti matematické lingvistiky; National Bureau of Standards sponzoroval symposium s názvem "Statistical Association Methods for Mechanized Documentation." Ze sympozia

vzešlo několik důležitých textů, včetně první zmínky o SMART systému G. Saltona.
    - polovina 60. let: National Library of Medicine vyvinula MEDLARS Medical Literature Analysis and Retrieval System, první velkou strojově čitelnou databázi a batch-retrieval systém; projekt Intrex na MIT.
    - 1965: J. C. R. Licklider publikuje Libraries of the Future.
    - 1966: Don Swanson se zapojil do studia požadavků na budoucí katalogy na University of Chicago
    - konec 60. let: F. Wilfrid Lancaster dokončil hodnotící studie systému Medlars a publikoval první verzi svého textu týkajícího se získávání informací.
    - 1968 : Gerard Salton zveřejnil Automatic Information Organization and Retrieval; John W. Sammon, Jr. nastínil ve své zprávě "Some Mathematics of Information Storage and Retrieval..."

vektorový model

- - 1969: Sammonův text "A nonlinear mapping for data structure analysis" (IEEE Transactions on Computers) byl prvním návrhem vizuálního rozhraní pro systém získávání informací.
- 70. léta
    - počátek 70. let: První online systémy—NLM's AIM-TWX, MEDLINE; Lockheed's Dialog; SDC's ORBIT; Theodor Nelson představuje koncept hypertextu, je publikována jeho kniha Computer Lib/Dream Machines.
    - 1971: Nicholas Jardine a Cornelis J. van Rijsbergen zveřejňují text "The use of hierarchic clustering in information retrieval", ve kterém vyslovují myšlenku "cluster hypothesis."
    - 1978: První ročník ACM SIGIR konference.
    - 1979: C. J. van Rijsbergen publikoval text Information Retrieval. Silný důraz na modely pravděpodobnosti.
    - 1979: Tamas Doszkocs implementoval uživatelské rozhraní jazyka CITE pro systém MEDLINE v National Library of Medicine. CITE podporoval zadání volného formuláře, hodnocení výstupu a relevantní zpětnou vazbu.

- 80. léta
  - 1980: První ročník mezinárodní ACM SIGIR konference, pořádanou společně s British Computer Society IR group v Cambridge.
  - 1982: Nicholas J. Belkin, Robert N. Oddy, and Helen M. Brooks navrhli ASK (Anomalous State of Knowledge). To byla zásadní myšlenka, ačkoliv se jejich automatický analytický nástroj ukázal jako nedostačující.
  - 1983: Salton (a Michael J. McGill) vydali Introduction to Modern Information Retrieval, se silným důrazem na vektorové modely.
  - 1985: David Blair a Bill Maron vydávají: An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System
  - polovina 80. let: Snahy vyvinout komerční systémy získávání informací přímo pro koncové uživatele.
  - 1985–1993: Zásadní testy týkající se vizuálních rozhraní a experimentálních systémů téhož. Práce Donalda B. Crouche, Roberta R. Korfhageho, Matthewa Chalmerse, Anselma Spoerriho a dalších.
  - 1989: První návrhy Tima Bernerse-Lee k World Wide Web v CERN.
- 90. léta
  - 1992: První TREC konference.
  - 1997: Vydání knihy Information Storage and Retrieval Roberta F. Korfhageho s důrazem na vizualizaci a systémy multirefenčních bodů.
  - konec 90. let: Webové vyhledávače obsahují mnoho funkcí, které se dříve využívaly jen v experimentálních verzích systémů získávání dat. Vyhledávače se stali nejběžnějším a možná nejlepším využitím modelů získávání informací.

Reference

V tomto článku byl použit překlad textu z článku Information retrieval na anglické Wikipedii.

1. ↑ JANSEN, B. J.; RIEH, S. The Seventeen Theoretical Constructs of Information Searching and Information Retrieval. *Journal of the American Society for Information Sciences and Technology*. 2010, roč. 61, čís. 8, s. 1517–1534. Dostupné v archivu pořízeném dne 2016-03-04. (angličtina)

2. ↑ SANDERSON, Mark; CROFT, Bruce W. *The History of Information Retrieval Research* [online]. 2012. Dostupné online. (angličtina)

Externí odkazy

- Obrázky, zvuky či videa k tématu získávání informací na Wikimedia Commons