اینترنت و اطلاعات دولتی

رضا اردلان

چکیده

سیستم توماس به گونه‌ای طراحی شده است تا بتوان به کمک آن اطلاعات دولتی را بر روی اینترنت‌در دسترس عموم قرار دارد. این نمونه می‌تواند برای کتابخانه‌های دیجیتالی دولتی به کار رود. طرح‌توماس پروژه مشترک کتابخانه کنگره آمریکا و دانشگاه ماساچوست است که فرصت مهمی در بازیابی و تکنیک‌های واسط کاربر (User Interface) ایجاد می‌کند که برای دسترسی موثر به اطلاعات پیچیده‌در محیط‌های اینترنتی ضروری است. تجربیات اولیه در استفاده از توماس نشان داده است که نیاز زیادی‌به بازیابی این گونه اطلاعات وجود دارد و پرسش‌های کاربر بایستی کوتاه‌تر از نمونه‌های قبلی مانندTREC باشد. تکنیک‌هایی مانند پردازش پرسش، گسترش پرسش و پردازش واژگانی، همگی باید به‌صورت‌همکاری‌های چند جانبه پیگیری شود و ارتقا یابد. کلید واژه‌های این بررسی عبارتند از: بازیابی‌اطلاعات، اینترنت، پردازش پرسش.

مقدمه

در اواسط دسامبر ۱۹۹۴ رییس جدید مجلس نمایندگان کنگره آمریکا از کتابخانه کنگره خواست تایک سیستم جدید اطلاعاتی برای ارائه اطلاعات برروی اینترنت ساخته شود. این پایگاه اطلاعاتی‌جدید می‌بایست کانون توزیع اطلاعات مربوط به کلیه فعالیت‌های کنگره، شامل مشروح مذاکرات‌مجلس، مصوبات مجلس، کلیه اسناد مجلس نمایندگان و سنا، آدرس‌های پست الکترونیکی و ارتباط(LINK) با سایر منابع الکترونیکی قانونگذاری موجود در اینترنت باشد.

کتابخانه کنگره آمریکا به عنوان سایت مرکزی این گروه از اطلاعات در نظر گرفته شد. البته از قبل، بخشی یا تمام اطلاعات موجود با استفاده از امکانات گوفر و تل نت در کتابخانه سنا، مجلس نمایندگان،اداره امور چاپی دولتی، و انواع مختلف خدمات تجاری قابل جستجو بود. رییس جدید مجلس‌نمایندگان از کتابخانه کنگره خواسته بود تا این سیستم جدید بتواند عموم مردم را به آسانی به WWWمرتبط سازد.

پروژه توماس طرح مشترک کتابخانه کنگره و دانشگاه MIT آمریکاست. در این طرح مشترک علاوه برکتابخانه و دانشگاه ایالت ماساچوست، شرکایی از بخش‌های دولتی و تجاری حضور دارند.

بخش اطلاعات هوشمند دانشگاه ماساچوست آمریکا برنامه نرم‌افزاری این سیستم را تهیه کرد وبرای ایجاد تغییرات، قابلیت هایی را در این سیستم به وجود آورد. کتابخانه کنگره آمریکا نیز برای ارتباط‌ساده کاربران با برنامه، صفحات میانجی و امکان انتقال پایگاه‌های اطلاعات به HTML را تهیه کرد.

در کتابخانه کنگره آمریکا اسناد و اطلاعاتی همچون عکس‌های مربوط به جنگ جهانی دوم، تصاویرمتحرک قدیمی، کاست‌های صدا، اطلاعات مربوط به جنگ ویتنام، مطالعات کشور مربوط به بخش‌فدرال وجود دارد. این مواد را بایستی در زمره اطلاعات دیجیتال به حساب آورد. به همین دلیل کارشناسان اعتقاد دارند که طرح توماس طرح اولیه کتابخانه دیجیتال مبتنی بر اطلاعات دولتی است.

از ویژگی هایی یک کتابخانه دیجیتال می‌توان به فراهم آوردن امکان دستیابی به بخش گسترده‌ای از اطلاعات ارزشمند در سطح شبکه اشاره کرد که این طرح این قابلیت را داراست. دسترسی رایگان‌اطلاعات کتابخانه کنگره موجب شده است که طرح توماس از جهاتی نیز شبیه به کتابخانه‌های عمومی‌باشد.

پایگاه اطلاعاتی توماس

این پایگاه اطلاعاتی مختص اطلاعات قانون گذاری است. در زمان حاضر در این پایگاه می‌توان به اطلاعات مجلس ۱۰۳ و ۱۰۴ کنگره و متون مربوط به مذاکرات این دو دوره که در ۱۰ جلد کتاب چاپ‌شده است، اشاره کرد.

مشروح مذاکرات مجلس سنا و مجلس نمایندگان (کنگره آمریکا) به صورت روزانه با تشکیل حداقل‌یک جلسه به چاپ می‌رسد. هر رکورد در این پایگاه شامل Daily Digest است که شامل خلاصه رویدادها به صورت روزانه و یک بخش از مجلس نمایندگان و یک بخش از مجلس سنا و بخش مشروح‌مذاکرات است. بخش مشروح مذاکرات شامل اظهارات نمایندگان است که در صحن علنی ایراد نشده‌است، اما در پایگاه موجود است. اطلاعات بخش‌های سنا و نمایندگان به مباحث یا مذاکراتی که درباره‌موضوعات خاص ایراد گردیده است، تقسیم می‌شود که هر کدام دارای عناوین خاص است؛ مانند"اصلاحیه تنظیم بودجه". این بخش‌ها در مجموع سندهای این پایگاه را تشکیل می‌دهد. هر یک از این‌لوایح دارای حجم مشخصی است از ۱ کیلو بایت گرفته تا ۷۰۰ کیلوبایت. به طور مثال حجم اطلاعات مربوط به نمایندگان کنگره در جلسه صد و سوم در حدود ۳۰۰۰۰ سند است که اندازه هر یک ۶۰۰مگابایت است.

اطلاعات مجلسین به صورت روزانه از طرف اداره اطلاعات چاپی دولتی با فرمت FTP ارسال‌می‌گردد. پس از دریافت اطلاعات اولیه، به هر سند برچسب عنوان داده می‌شود. سپس کدهای اداره اطلاعات چاپی به کدهای HTML تبدیل می‌شود و برای هر مورد فهرست مندرجات ساخته می‌شود.

علاوه بر این موارد، یک فهرست مندرجات فرامتنی برای پیشینه‌های مجلسین ایجاد می‌گردد. پس ازاتمام این مراحل، اطلاعات مورد نظر نمایه سازی می‌شود. لوایح براساس عنوان، شماره لایحه، و متن‌لایحه، نمایه سازی می‌شوند و چون سرعت نمایه سازی بالا است، می‌توان اطلاعات را در کمتر از یک‌روز نمایه سازی کرد.

سیستم بازیابی اطلاعات در طرح توماس

موتور جستجوی مورد استفاده در طرح توماس براساس مدل احتمالات بازیابی اطلاعات است. از این‌سیستم در بسیاری از پروژه‌های پژوهشی استفاده شده و در بازیابی اطلاعات دولتی کاربردی موثرداشته است. از ویژگی‌های این سیستم می‌توان به موارد زیر اشاره کرد:

۱- ستاده‌های دسته بندی شده

احتمال ارتباط بین متن و پرسش (Query) از طریق ترکیب متشابهات موجود در متن با کل اطلاعات. کارایی این تکنیک از طریق روش دقت / بازیابی تایید شده است.

۲- بازیابی براساس متن

احتمال مرتبط بودن یک سند هم براساس کل مندرجات یک سند و نیز بهترین متن همانند در سنداست. این تکنیک کارایی بازیابی را بالا می‌برد و ابزار بررسی تمام سند را فراهم می‌سازد. از این تکنیک به این دلیل در طرح توماس استفاده شده است که می‌تواند در لوایح پرحجم، ابزاری برای جلوگیری از جستجوی بخش به بخش باشد.

۳- توانایی انجام پرسش‌های ساده و پیچیده

زبان بازیابی اطلاعات در این تکنیک می‌تواند عملگرهایی را فعال سازد که تعیین می‌کند چگونه بایستی نشانه‌ها در داخل سند برای رسیدن به ربط احتمالی ترکیب شوند. این بدان معنی است که قالب احتمال‌گرای بازیابی را می‌توان برای پرسش‌های تک واژه، ترکیبات بول، عبارتی، یا هر ترکیب دیگری به‌کار برد.

۴- بازیابی مبتنی بر حیطه (Field)

جستجوگر می‌تواند براساس فیلد نیز به نمایه سازی و بازیابی اطلاعات بپردازد. به عبارتی گسترش‌فیلد می‌تواند بخشی از ارزشیابی احتمالی پرسش باشد. بدین معنا که می‌توان جستجو با فیلد را باپرسش‌های پیچیده ترکیب کرد تا به درجه بندی اطلاعات دست یافت. در طرح توماس از این تکنیک‌می‌توان برای بازیابی برخی از استنادات سند نظیر شماره لایحه و نوع (Type) آن استفاده کرد.

۵- نمایه سازی انعطاف‌پذیر و موثر

روند نمایه سازی در طرح توماس به گونه‌ای است که می‌تواند برای هر ساختاری اعم از:... MARC, HTMEL، تکنیک‌های واژه‌شناسی (تکنیک‌های ریشه‌شناسی) و تشخیص گرهای مفهومی حوزه (Domain) مانند بسامد اشخاص، شرکت‌ها، مکان‌ها، و تاریخ‌ها به کار رود. در حال حاضر از تشخیص گرها در توماس استفاده نمی‌شود، زیرا نمی‌توانیم پیش بینی کنیم که چه مفاهیمی دراین طرح ارزشمند هستند.

۶- ابزارهای پردازش پرسش و گسترش پرسش

پرسش‌های زبان طبیعی می‌تواند به استعلام پرسش‌ها تبدیل شود و از ابزارهایی نظیر برچسب نقش‌گرامری و تشخیص گر اتمام جمله استفاده کند. پرسش‌ها قابلیت گسترش خودکار دارند و از عبارت‌های مرتبط متن استفاده کنند.

۷- پشتیبانی از ربط بازخورد و مسیر دهی

بازخورد کاربر در مورد ربط سندهای بازیابی شده می‌تواند به صورت خودکار نمایانگر پرسشی دریک جلسه یا اطلاعات پرسنلی اشخاص در محیطی باشدکه اطلاعات دریافتی با اطلاعات ذخیره شده مقایسه می‌شود.

کارایی سیستم

یکی از راه‌های بررسی کارایی سیستم توماس، توجه به تعداد نتایج جستجو است که این سیستم انجام‌داده است. در بررسی یک دوره زمانی سه ماهه در سال ۱۹۹۵ تعداد ۲۳۰۲۵۸۹ تراکنش اطلاعاتی‌وجود داشته است که موجب ۲۹۴۵۷۵ دسترسی به صفحات اینترنتی توماس بوده است. از میزان‌دسترسی‌ها می‌توان حدث زد که تعداد ۹۴۹۱۱ پرسش وجود داشته است.

نمونهٔ زیر را می‌توان برای آشنایی بیشتر محققان ارائه داد:

Query Count	
balanced budget	2/600
crime	1/057
gun(s	994
balanced budget amendment	991
s	314 902
telecommunications	888
welfare	846
budget	753
abortion	678
line item veto	610
gun control	539
unfunded mandates	532
welfare reform	513
education	441
tax	415
term limits	401
crime bill	375
contract with America	366
public broadcasting	333
decency	333
immigration	316
balanced	315
health care	305
baseball	303
firearms	300
TOTAL	16/106

مشکلات برنامه توماس

در ژانویه سال ۱۹۹۵ یکی از متخصصان، مقاله‌ای جنجالی درباره توماس نوشت و در مقاله‌اش ادعاکرد که، «تکنیک‌های جستجو در این برنامه کارایی لازم را ندارد». نویسنده این مقاله، عبارت "Elderly black Americans" به معنای سیاه پوستان سالمند آمریکا را در سیستم وارد کرده بود ولایحه‌ای درباره "خرس‌های سیاه" دریافت کرده بود در ادامه همین جستجو، دانشگاه‌ها و کالج‌های‌مربوط به سیاه پوستان نیز ذکر شده بود. علت این اشتباه این گونه بود که چون در سیستم هیچ گونه اطلاعاتی درباره سیاه پوستان سالمند وجود نداشت، ناخواسته واژه‌های سیاه پوستان و آمریکا بازیابی‌شده بود. در حال حاضر ما سرگرم مطالعه در مورد احتمال افزودن یک تزاروس از پیش آماده شده به‌سیستم توماس یا احتمال استفاده از یک تزاروس خودکار مانند "عبارت یاب" هستیم.

استفاده از ریشه یابی خودکار توماس جنبه‌های مثبت و منفی داشته است. ریشه یابی روش خوبی‌برای بازیابی واژه‌های مختلفی است که یک گروه از مفاهیم مجرد را تشکیل می‌دهند.

نتیجه‌گیری

پردازش اطلاعات که در بالا به آن اشاره شد، چون نمونه قبلی نداشته است، با روش آزمایش و خطابه دست آمده است. تحلیل گران سیستم توماس به طور مستمر در حال تعیین کارایی پرسش‌های کاربران‌و بررسی تکنیک‌های مورد استفاده هستند. تا به حال بر اساس اطلاعاتی که از استفاده کنندگان به دست‌آمده است، این سیستم موفق عمل کرده است.

اکنون آزمایش‌های بیشتری براساس ربط رسمی قضاوت‌ها برای تعیین سطح بهبود کارایی‌تکنیک‌های جدید به دست آمده است.

REFERENCES

1. J.P. Callan. Passage-level evidence in document retrieval.In Proceedings of ACM SIGIR International Conference on Research and Development in Infornation Retrieval, pages 302-310, 1994.

2. W.B. Croft and J. Xu. Corpus -specific stemming using word from co-occurrence. In Fourth Annual Symposium on Document Analysis and Information Retrieval, pages 147-159, 1995.

3. D. Harman. Overview ofthe Third Text Retrieval Conference (TREC-3). In D. Harman, editor, Proceedings of the Third Text Retrieval Conference (TREC-3), pages 1-120. NIST Special Publication 500-225, 1995.

4. Y. Jing and W.B. Croft.An association thesaurus for information retrieval. In Proceedings of RIAO 94, pages 146-160, 1994.

5. Robert Krovetz. Viewing morphology as an inference process. In Proceedings of the 16th International Conference on Research and Development in Information Retrieval, pages 191-202, 1993.

6. M. Putzel. Room for doubting Thomas. Boston Globe, page 92, January 27, 1995.

7. T.B. Rajashekar and W.B. Croft. Combining automatic andmanual index representations in probabilistic retrieval. Journal of the American Society for Information Science, 46 (4): 272-283, 1995.

8. Gerard Salton and Michael J. McGill. Introduction to Modern Information Retrieval. McGraw-Hill, 1983.

9. Karen Sparck Jones, editor. Information Retrieval Experiment. Butterworth,1981.

10. Howard Turtle. Natural language vs. Boolean query evaluation: A comparison of retrieval performance. In Proceedings ACM SIGIR 94, pages 212-220, 1994.

11. H.R. Turtle and W.B. Croft. Evaluation of an inference network-based retrieval model. ACM Transactions on Informations Systems, 9 (3): 187-222, 1991.

12. H.R. Turtle and W.B. Croft. 1992. A comparsion of text retrieval models. Computer Journal, 35 (3): 279-290, 1992.