University of Wollongong
Research Online

Faculty of Commerce - Papers (Archive) Faculty of Business and Law

1-1-2005

The evaluation of latent semantic indexing in Persian text retrieval

Farhad Oroumchian
University of Wollongong, farhado@uow.edu.au

H. Bashiri

M. Rohani

A. Moeini

Follow this and additional works at: https://ro.uow.edu.au/commpapers

Part of the Business Commons, and the Social and Behavioral Sciences Commons

Recommended Citation

Oroumchian, Farhad; Bashiri, H.; Rohani, M.; and Moeini, A.: The evaluation of latent semantic indexing in Persian text retrieval 2005, 150-156.
https://ro.uow.edu.au/commpapers/2386

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: research-pubs@uow.edu.au

The evaluation of latent semantic indexing in Persian text retrieval

Keywords

Evaluation, Latent, Semantic, Indexing, Persian, Text, Retrieval

Disciplines

Business | Social and Behavioral Sciences

Publication Details

Oroumchian, F., Bashiri, H., Rohani, M. & Moeini, A. (2005). The evaluation of latent semantic indexing in Persian text retrieval. In A. Khademzadeh, S. Shahnazi & K. Badie (Eds.), Proceedings of the 10th Annual Computer Society of Iran Computer Conference CSCIC 2005 (pp. 150-156). Tehran, Iran: Sadegh.

This conference paper is available at Research Online: https://ro.uow.edu.au/commpapers/2386

دهمین
کنفرانس سالانه انجمن کامپیوتر ایران
۲۷ تا ۲۹ بهمن ۱۳۸۳

ارزیابی مدل نمایه‌سازی معنایی پنهان در بازیابی متون فارسی

حسن بشیری
دانشجوی کارشناسی ارشد نرم‌افزار
دانشگاه شهید بهشتی

دکتر فرهاد ارومچیان
عضو هیات علمی دانشکده‌ی فنی
دانشگاه تهران

سید محمدتقی روحانی رانکوهی
عضو هیات علمی دانشکده‌ی برق و کامپیوتر
دانشگاه شهید بهشتی

دکتر علی معینی
عضو هیات علمی دانشکده‌ی فنی
دانشگاه تهران

hbashiri, foroumchian@acm.org

چکیده

امروزه اطلاعات از چنان درجه‌ی اهمیتی برخوردار است که عصر حاضر را عصر اطلاعات نامیده‌اند و میزان اطلاعات تولید شده و میزان استفاده از اطلاعات دو معیار اساسی برای توسعه کشور به شمار می‌آید. اما تولید و وجود اطلاعات کافی نیست بلکه باید ابزارهایی را برای تسهیل در استفاده از این اطلاعات برای کاربران فراهم کرد. در واقع سوال اصلی کاربران این است که چگونه باید به نیاز اطلاعاتی خود در این حجم عظیم منابع اطلاعاتی پاسخ دهند. سیستم‌های بازیابی اطلاعات مهم‌ترین ابزار، برای پاسخ به نیاز اطلاعاتی کاربران هستند که امروزه بیش از پیش مورد توجه قرار گرفته‌اند. مدل‌های مختلف پیاده‌سازی شده در این سیستم‌ها، در زبان‌های مختلف نتایج متفاوتی را به همراه دارد. مدل فضای برداری از جمله پایه‌ای ترین مدل‌ها، در این دسته از سیستم‌ها است. اما بیشتر مدل‌های بازیابی اطلاعات بر حضور یا عدم حضور واژه به کار رفته در متن تاکید دارند. در این بین مدل نمایه‌سازی معنایی پنهان از مدل‌های مفهومی در سیستم‌های بازیابی اطلاعات است که به دنبال تطبیق مفهومی پرسش کاربر با اسناد مجموعه است. در این مقاله، این مدل با تکیه بر تفکیک ماتریس به روش تجزیه نیمه جدا، پیاده‌سازی و ارزیابی شده است.

کلمات کلیدی

بازیابی اطلاعات، نمایه‌سازی معنایی پنهان، تجزیه نیمه جدا ماتریس و مجموعه آزمایش

۱. مقدمه

بازیابی اطلاعات[1] تکنیک‌های نمایش، ذخیره‌سازی، سازماندهی و دسترسی به فقره‌های اطلاعاتی است که به منظور پاسخ به نیاز اطلاعاتی کاربران مورد استفاده قرار می‌گیرد. تشخیص نیاز اطلاعاتی کاربر، بازیابی اطلاع یا اسناد مرتبط به پرسش و نمایش اسناد بازیابی شده در یک رتبه‌بندی مناسب از مهم‌ترین فعالیت‌های سیستم بازیابی اطلاعات است. مهم‌ترین چالش این سیستم‌های کار با حجم بالایی از اسناد و اطلاعات است که محققان را برای یافتن مدل بازیابی با کارایی و زمان بازیابی مناسب تشویق می‌کند.

[1] Information Retrieval

۲. مدل فضای برداری

مدل فضای برداری پایه‌ای‌ترین مدل در سیستم‌های بازیابی اطلاعات است که توسط Salton ابداع شد [۱]. در این مدل ابتدا سند به برداری تبدیل می‌شود که حاوی کلمات مهم متن سند، به همراه وزن هر کلمه بر اساس میزان تاثیرگذاری کلمه بر محتوی متن در مقایسه با سایر کلمات است. تهیه بردار برای هر سند بر اساس تکنیکی به نام نمایه‌سازی صورت می‌گیرد. در نمایه‌سازی ابتدا کلمات عمومی از متن حذف می‌گردند و کلمات باقیمانده ریشه‌یابی می‌شوند. سپس بر اساس پارامترهای مختلفی مانند تعداد تکرار کلمه در متن، تعداد تکرار کلمه در اسناد مجموعه و مولفه‌های نرمال‌سازی وزنی به هر کلمه نسبت داده می‌شود [۲ و ۳]. همین فعالیت‌ها برای پرسش کاربر نیز تکرار می‌شود. به این ترتیب هر سند از مجموعه ای از کلمات به برداری تبدیل می‌شود که در فضای جدیدی به نام فضای برداری قرار دارد. در این فضا که بسته به تعداد کلمات مجموعه یک فضای n بعدی است، بردار هر سند ترسیم می‌شود. پرسش کاربر نیز بعد از اعمال فعالیت‌های نمایه‌سازی به برداری تبدیل می‌شود که در فضای جدید ترسیم می‌گردد. در این فضا هر سندی که به پرسش کاربر نزدیک‌تر باشد سند مرتبط شناخته می‌شود و بازیابی می‌گردد. معیار نزدیکی در این فضا زاویه‌ای است که بردار پرسش با هر یک از بردارهای سند می‌سازد. این میزان نزدیکی، معمولا با رابطه زیر که به نام مشابهت کسینوسی شناخته می‌شود، محاسبه می‌گردد [۱ و ۲]:

$$Sim(q_i, d_j) = \frac{\overrightarrow{q_i} \bullet \overrightarrow{d_j}}{|\overrightarrow{q_i}| \times |\overrightarrow{d_j}|} = \frac{\sum_{k=1}^{t} w_{ki} \times w_{kj}}{\sqrt{\sum_{k=1}^{t} w_{ki}^2} \times \sqrt{\sum_{k=1}^{t} w_{kj}^2}}$$

در این رابطه q_i بردار پرسش کاربر، d_j بردار سند jام، w_{ki} وزن کلمه‌ی kام در پرسش کاربر و w_{kj} وزن کلمه kام در سند d_j است.

۳. مدل نمایه‌سازی معنایی پنهان

مدل فضای برداری، مدلی انعطاف‌پذیر و پارامتریک در بازیابی اطلاعات است که با اعمال روش‌های مختلف وزن‌دهی کلمات و استفاده از سایر صور تشخیص نزدیکی به سند مانند مشابهت ژاکارد می‌تواند نتایج متفاوتی را بدست آورد. اما آنچه در این مدل مورد توجه قرار می‌گیرد حضور کلمه در متن است. در واقع این مدل از انطباق واژه‌های به‌کار رفته در پرسش کاربر با واژه‌های اسناد استفاده می‌کند. به این ترتیب استفاده از اصطلاح "**مدل بازیابی اطلاعات مبتنی بر واژه**" برای این دسته از مدل‌ها که در فضای برداری عمل می‌کنند توصیف مناسبی از عملکرد آنها است [۲]. اگر چه چنین

روش‌هایی در سیستم‌های بازیابی اطلاعات توانسته‌اند بسیاری از مشکلات مربوط به چگونگی بازیابی یک سند مرتبط را حل کند اما از آنجائیکه کلمات ممکن است معنای متفاوت داشته باشند، این مدل‌ها نمی‌توانند همیشه روش دقیق و مناسبی برای بازیابی اسناد مرتبط باشند. مشکل اصلی از آنجا ناشی می‌شود که کاربران با ارائه پرسش به دنبال اسنادی هستند که از نظر مفهومی به نیاز اطلاعاتی آنها نزدیک است. اما با استفاده از مدل‌های موجود که بر اساس اشتراک واژه‌ها بین پرسش کاربر و سند کار می‌کنند، تنها اسنادی بازیابی خواهند شد که واژه‌ی مطرح شده در پرسش کاربر را داشته باشند و حال آنکه ممکن است سندی از نظر مفهومی با پرسش کاربر مرتبط باشد اما به دلیل عدم اشتراک واژه‌ای بازیابی نشود. این مشکل از داشتن چند مترادف[1] و گاه چند معنی متفاوت برای یک واژه[2] ناشی می‌شود. کاربران در زمینه‌های مختلف یا با نیازهای متفاوت، دانش یا عادت‌های زبانی، اطلاعات یکسانی را با واژه‌های متفاوت بیان می‌کنند. گستردگی مترادف کلمات باعث کاهش معیار بازیابی[3] در سیستم‌های بازیابی اطلاعات می‌شود. به همین ترتیب وجود چند معنی متفاوت و مجزا برای هر کلمه معیار دقت[4] را در سیستم‌های بازیابی اطلاعات کاهش می‌دهد. روش پیشنهادی نمایه‌سازی معنایی پنهان[5] (LSI) متن، روشی برای غلبه بر مشکل انطباق واژه‌هاست.

در این روش فرض می‌شود ساختار پنهانی در متن وجود دارد که می‌تواند در بازیابی صحیح و انطباق مفهومی سند با پرسش کاربر موثر باشد [٤، ٥ و ٦]. در LSI برای کشف این ساختار پنهان از تکنیک‌های آماری و جبر خطی استفاده می‌شود.

۳-۱: روش تجزیه نیمه جدا ماتریس

روش‌های تفکیک ماتریس، ایده کشف ساختار پنهان و تغییر فضا در ماتریس‌های دو بعدی است. در [٧] تحقیق جامعی پیرامون روش‌های تفکیک ماتریس شده است. دو روش تجزیه تک مقداری[6] (SVD) و تجزیه نیمه جدا[7] (SDD) ماتریس از مهم‌ترین روش‌های تفکیک ماتریس هستند. به دلیل حجم بالای محاسبات در روش SVD ما در پیاده‌سازی مدل LSI از روش SDD استفاده کرده‌ایم.

برای نگهداری اطلاعاتی در مورد واژه‌ها، حضور یا عدم حضور آنها در اسناد و وزن متناظر واژه در هر سند، از ماتریس Term×Document استفاده می‌شود. عمدتا این ماتریس برای هر مجموعه اسناد

---

[1] Synonym
[2] Polysemy
[3] Recall
[4] Precision
[5] Latent Semantic Indexing
[6] Singular Value Decomposition
[7] Semi-Discrete Decomposition

یک ماتریس خلوت[1] است. مطابق آمار ارائه شده در [٦] تنها ۱٪ از درایه‌های چنین ماتریسی غیر صفر است. به همین دلیل ایده‌ی کاهش فضا می‌تواند بر کارائی سیستم بازیابی تاثیر به‌سزایی داشته باشد. روش SDD تکنیکی برای کاهش فضای ماتریس Term×Document است. نکته‌ی جالب این روش کاهش خطای ناشی از تنوع در مترادف و معانی کلمات به همراه کاهش فضای ماتریس Term×Document است.

روش SDD بر پایه تئوری جبر خطی است. در این روش برای هر ماتریس $A_{m\times n}$ که در آن تعداد ردیف‌های ماتریس (m) بزرگتر یا مساوی تعداد ستون‌های ماتریس (n) باشد ($m \geq n$) می‌توان ماتریس A را بصورت حاصل‌ضرب یک ماتریس متعامد ستونی $X_{m\times n}$، ماتریس قطری $D_{n\times n}$ با درایه‌های مثبت یا صفر (مقادیر منفرد) و ترانهاده ماتریس متعامد $Y_{n\times n}$ نوشت. تجزیه‌ی تک مقدار ماتریس A با رابطه زیر تعریف می‌شود [۸ و ۹].

$$\begin{pmatrix} & & \\ & A & \\ & & \end{pmatrix}_{m\times n} = \begin{pmatrix} & & \\ & X & \\ & & \end{pmatrix}_{m\times n} \cdot \begin{pmatrix} d_1 & & \\ & \cdots & \\ & & d_n \end{pmatrix}_{n\times n} \cdot \begin{pmatrix} & & \\ & Y^T & \\ & & \end{pmatrix}_{n\times n}$$

در این رابطه $X^TX = Y^TY = I$ و $D = diag(d_1, d_2, \ldots, d_n)$ که برای $1 \leq i \leq r$ ، $d_i>0$ و برای $d_j = 0 , j \geq r + 1$. در این رابطه r ستون نخست از ماتریس‌های متعامد X و Y از روی بردارهای ویژه متعامد و متناظر با r بردار ویژه غیرصفر از AA^T و A^TA به ترتیب بدست می‌آیند.

۳-۲: بازیابی در مدل LSI

مشابهت پرسش کاربر به سند در این مدل با استفاده از رابطه $S = \tilde{q}^T \tilde{A}$ بدست می‌آید. در این رابطه بردار A و بردار پرسش کاربر در فضای k بعدی جدید نگاشت می‌شود که این نگاشت با استفاده از رابطه‌های زیر انجام می‌شود [۸ و ۹] .

$$\tilde{A} = D_k^{1-\alpha} Y_k^T$$
$$\tilde{q} = D_k^{\alpha} X_k^T q$$

در این رابطه α پارامتر شکستن است که بصورت پیش فرض مقدار صفر برای آن در نظر گرفته شده است. در آزمایشات مختلف مقدار α=0.5 نیز نتیجه دقت میانگین را افزایش داده است.

[1] Sparse Matrix

روش وزن‌دهی به کار گرفته شده در این آزمایش روش Lnu برای وزن‌دهی به واژه‌های اسناد و ltu برای وزن‌دهی به واژه‌های پرسش کاربر است. رابطه وزن‌دهی به واژه‌های سند مطابق با رابطه زیر محاسبه شده است [۱۰].

$$wd_{ij} = \frac{\dfrac{1+\log(tf_i)}{1+\log(averagetf_j)}}{(slope \times NUT_j) + (1-slope) \times Pivot}$$

پارامتر wd_{ij} وزن کلمه iام از سند jام را نشان می‌دهد. در این رابطه $L = \dfrac{1+\log(tf_i)}{1+\log(averagetf_j)}$ است که tf تعداد تکرار کلمه در سند را نشان می‌دهد. پارامتر average tf نیز میانگین تعداد تکرار کلمات برای یک سند را مشخص می‌کند. در واقع با تقسیم tf بر average tf وزن کلمه در سند نرمال می‌شود. پارامتر n در رابطه Lnu یک پارامتر دودویی است که نشان می‌دهد که تعداد اسنادی که حاوی کلمه هستند در نظر گرفته شود یا خیر. معمولا در بیشتر آزمایشات n=1 در نظر گرفته می‌شود. همچنین در این رابطه $u = \dfrac{1}{(slope \times NUT_j) + (1-slope) \times Pivot}$ است که در آن slope و Pivot پارامترهای مجموعه هستند که slope شیب منحنی فضای مجموعه را مشخص می‌کند و معمولا مقدار slope=0.25 برای بیشتر مجموعه‌ها مقدار پیشنهادی است. پارامتر Pivot هم میانگین طول اسناد موجود در مجموعه در نظر گرفته شده است که برای مجموعه مورد ارزیابی ما این مقدار به 35.2735 تنظیم شده است. پارامتر سوم در این رابطه NUT [1] است که تعداد کلمات با رخداد tf=1 را در سند مشخص می‌کند. همچنین برای وزن‌دهی به کلمات پرسش نیز از رابطه ltu استفاده شده است که مطابق با رابطه زیر تعریف می‌شود [۱۰].

$$\frac{(\ln(tf)+1.0) \times \ln\dfrac{N}{n}}{(slope \times NUT) + (1-slope) \times Pivot}$$

۳-۳: مجموعه آزمایش

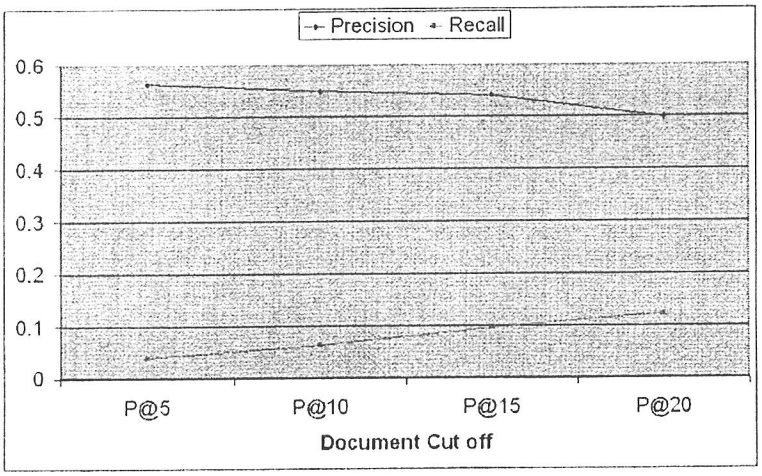
این مجموعه، که مجموعه‌ای منحصر به‌فرد در ارزیابی مدل‌های بازیابی روی متون فارسی است، شامل قوانین ۹۰ سال کشور ایران است. در این مجموعه، اسناد از نظر طول دارای تنوع زیادی هستند. برای مثال یک قانون کوچک با یک یا چند پاراگراف به عنوان یک سند مورد توجه قرار گرفته است و همچنین کل بودجه‌ی سالیانه‌ی کشور با همه‌ی بخش‌ها، زیربخش‌ها و تبصره‌ها نیز به عنوان یک سند

[1] Number of Unique Terms

بررسی شده است. برای بررسی مدل‌های بازیابی، مستندات موجود در مجموعه قوانین به قطعات کوچکتری[1] تقسیم شده‌اند که هر قطعه شامل یک بخش یا زیربخش از یک قانون است و به این ترتیب هر سند بیش از چند پاراگراف نیست. با این تقسیم‌بندی ۱۷۷۰۸۹ قطعه در مجموعه ایجاد شده است.
در این مجموعه یک گروه از وکلا بر تمام ورودی‌ها و عملگرهای سیستم نظارت کرده‌اند. برای ارزیابی این سیستم ٤١ پرسش مورد استفاده قرار گرفته است و به ازاء هر پرسش، ۲۰ سند بازیابی شده‌ی اول، داوری شده‌اند. برای هر سند امتیازی بین ٤-۰ توسط داور انسانی در نظر گرفته شده است. امتیاز صفر به معنی بی‌ارتباط بودن کامل یک سند به پرسش و امتیاز ٤ به معنی ارتباط کامل یک سند به پرسش است. تعداد واژه‌های موجود در این مجموعه ۷۸۸۹۵ کلمه است که در کنار تعداد اسناد آن یعنی ۱۷۷۰۸۹ سند، مجموعه‌ای بسیار بزرگ برای ارزیابی سیستم‌های بازیابی اطلاعات فارسی فراهم کرده است.

۳-۴: نتایج مدل

در ارزیابی مولفه پیاده‌سازی شده ۱۵ پرسش از مجموعه قوانین انتخاب شد. ماتریس Term×Document ورودی با مولفه SDDPACK با پارامتر k=200 به سه مولفه شکسته شد. معیار دقت و بازخوانی در برش‌های[2] ۵، ۱۰، ۱۵ و ۲۰ برای هر یک از پانزده پرسش ورودی محاسبه شد.
در نهایت به ازاء هر ۱۵ پرسش ارزیابی شده در برش‌های سندی ذکر شده میانگین دقت و بازخوانی محاسبه شد و در نمودار آن در شکل ۱ ترسیم شده است.



شکل ۱: نمودار Precision-Recall برای میانگین ۱۵ پرسش مجموعه قوانین

[1] Passage
[2] Cut-Off

۴. نتیجه‌گیری و ادامه کار

استفاده از روش‌های مختلف وزن‌دهی به کلمات متن و پرسش و همچنین استفاده از مولفه SVDPACK به جای SDDPACK از جمله آزمایشاتی است که می‌توان برای کسب نتایج بهتر انجام داد. مجموعه‌ای که در ارزیابی مدل LSI مورد استفاده قرار گرفت مجموعه‌ای منحصر به‌فرد برای متون فارسی است اما در مقایسه با سایر مجموعه‌های استاندارد ارزیابی، چگالی نامناسب و بسیار پائینی دارد. به ویژه پراکندگی بسیار زیاد ماتریس و چگالی بسیار کم آن کیفیت ارزیابی مدل را کاهش می‌دهد. در صورت تهیه مجموعه‌ای با چگالی بیشتر انتظار می‌رود مدل نمایه‌سازی معنایی پنهان نتایج بهتری را به همراه داشته باشد.

۵. مراجع

[1] - G. Salton and M. J. McGill. "**Introduction to Modern Information Retrieval**", McGraw-Hill Book Co., New York, 1983.

[2] - Baeza-Yates R., Ribeiro-Neto B., "**Modern Information Retrieval**", ACM Press, 1999

[3] - Ed. Greengrass, "**Information Retrieval: A Survey**", Nov. 2000

[4] - Berry M. W., Dumais S. T., Letsche T. A., "**Computational Methods for Intelligent Information Access**", Proceedings of Supercomputing, Sandi ago, Ca, 1995

[5] - Deerwester S., Dumais S. T., Furans G. W., Landauer T. K., Harshman R., "**Indexing by Latent Semantic Analysis**", Journal of American Society for Information Science, 41, 1990, pp. 391-407

[6] - Berry M. W., Dumais S. T., O'Brien G. W., "**Using Linear Algebra for Intelligent Information Retrieval**". SIAM Review 37(4), 2000

[7] - D.B. Skillicorn, S. M. McConnell, E.Y. Soong, "**Handbook of Data Mining Using Matrix Decompositions**", School of Computing, Queen's University, Kingeston Canada, August 2003

[8] - Kolda T. G., O'Leary D. P., "**Latent Semantic Indexing via a Semi-Discrete Matrix Decomposition**", The Mathematics of Information Coding, Extraction and Distribution, Springer-Verlag, 1999, pp. 73{80.

[9] - Kolda T. G., O'Leary D. P., "**Algorithm 805: Computation and uses of the Semi-Discrete Matrix Decomposition**", ACM Transactions on Mathematical Software 26(3), 2000

[10] - Singhal, A., Buckley, C., Mandar, M. "**Pivoted Document Language Normalization**", Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.21-29, 1996.