مروری بر نمایه‌سازی معانی پنهان: نظریه و کاربردها

محمدباقر دستغیب[1]

چکیده

نمایه‌سازی معانی پنهان روشی است که اطلاعات را در قالب معانی ذخیره می‌کند، و از رابطهٔ پنهان میان اصطلاحات و متن استفاده می‌نماید. در این روش معایب توجه به اصطلاح در یک سند به صورت منفرد، برطرف می‌گردد. در نظام بازیابی که به این روش فراهم می‌آید، به‌جای تطبیق لغوی اصطلاحات کلیدی، رابطهٔ معنایی میان اصطلاحات نیز مورد توجه قرار می‌گیرد. در این مقاله، نظریه و کاربردهای نمایه‌سازی معانی پنهان در نظام‌های بازیابی مورد بررسی قرار می‌گیرند.

کلیدواژه‌ها: بازیابی اطلاعات، نمایه‌سازی معانی پنهان، تجزیهٔ مقدار ویژه

مقدمه

در میان گونه‌های مختلف اطلاعات موجود در اینترنت، بیشتر اطلاعات، بخصوص اسناد و مدارک علمی، دارای قالب‌بندی متنی می‌باشند و بنابراین بازیابی اطلاعات متنی از اهمیت بسیاری برخوردار است (Kowalski, 1997).

برای آنکه بازیابی اسناد در اینترنت امکان‌پذیر باشد، باید روشی مناسب برای پیاده‌سازی، ذخیرهٔ اسناد و نمایه‌سازی انتخاب گردد. در میان روش‌های گوناگون پیاده‌سازی سند و درخواست، غالباً روش «فضای بُرداری»[2] مورد استفاده قرار می‌گیرد. در این روش، سند و درخواست به صورت بُردارهایی از فرکانس یا وزن اصطلاحات نمایه، پیاده‌سازی می‌گردند. در میان فرمول‌های کلاسیک وزن‌دهی اصطلاحات نمایه، ساده‌ترین فرمول، استفاده از فراوانی، و بسامد معکوس می‌باشد: W=TF*IDF

در این روش، وزن هر اصطلاح از ضرب دفعات تکرار اصطلاح در سند (TF)[3] در دفعات تکرار اصطلاح در کل اسناد (IDF) به‌دست می‌آید. برای آن‌که طول اسناد در وزن اصطلاحات، اثرگذار نباشد می‌توان وزن نهایی را با استفاده از طول سند، «نرمال» کرد. درنهایت هر سند به یک بُردار تبدیل خواهد شد؛ با مقایسهٔ بین‌بُردار درخواست و بردار سند، اسناد مرتبط[4] بازیابی می‌گردند. یکی از روش‌های معمول برای محاسبهٔ شباهت میان بردارها، محاسبهٔ زاویهٔ میان بردار درخواست و بردار سند است. هرچه زاویهٔ میان این دو بردار کمتر باشد، سند و درخواست، شبیه‌ترند (شکل شماره 1)(Salton, 1983). درنهایت پس از نمایه‌سازی و محاسبهٔ وزن برای تمامی اصطلاحات کلیدی سند، یک ماتریس به نام ماتریس اصطلاح ـ سند[5] به دست خواهد آمد. هر سطر از این ماتریس، بردار مشخصهٔ یکی از اسناد می‌باشد و هر مدخل از ماتریس، وزن‌های

---

۱. کارشناس ارشد مهندسی رایانه-کتابخانهٔ منطقه‌ای علوم و تکنولوژی شیراز Dstghaib@srlst.com

2 . Vector Space Model
3 . Term Frequency
4 . Relevant
5. Term-Document Matrix

محاسبه‌شده در اسناد است (شکل شماره 2)(Salton, 1983).

عموماً اطلاعات با تطبیق لفظی کلمات درخواست با اسناد، بازیابی می‌شود. اگرچه تطبیق لغوی ممکن است روشی نادقیق در تطبیق درخواست و سند باشد، ولی از این شیوه استفاده می‌شود. معمولاً روش‌های زیادی برای بیان مفاهیم با واژه‌های مترادف وجود دارد، از این رو در روش لغوی، کلمات موجود در درخواست ممکن است با کلمات اسناد مشابه، تطبیق نیابند. از سوی دیگر، اغلب کلمات دارای معانی گوناگون هستند؛ بنابراین کلمات موجود در درخواست کاربر ممکن است به صورت لغوی، با کلمات موجود در اسناد غیرمرتبط تطبیق داده شوند. روش بهتر آن است که در زمان بازیابی اطلاعات، مفهوم و معنای پایهٔ اسناد مورد استفاده قرار گیرد(Bery, Dumais & Shippy, 1995; Rosario, 2000).

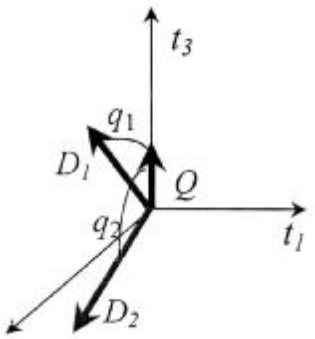
نمایه‌سازی معانی پنهان[6] بر آن است تا مشکل مقایسهٔ لغوی را با بهره‌گیری از شاخص‌های ادراکی آماری، و نه تکیه بر اصطلاحات به صورت منفرد، برطرف کند. نمایه‌سازی معانی پنهان بیان می‌کند که در به‌کارگیری اصطلاحات در متون، یک ساختار پنهان وجود دارد که با کاربرد و معنای اصطلاح در پاراگراف مرتبط است. برای آشکارسازی ساختار اصطلاح در سند، تجزیهٔ مقدار ویژه به کار برده می‌شود. سپس با استفاده از پایگاه دادهٔ مقادیر ویژه و بُردارهایی که از تجزیهٔ مقدار ویژه[7] برای اسناد به دست آمده، بازیابی انجام می‌شود. کارآیی نهایی سیستم بازیابی نشان می‌دهد بردارهایی که به صورت آماری از مقادیر ویژه به دست آمده‌اند، ارتباط معنایی اصطلاحات و اسناد را به شکلی نیرومندتر آشکار کرده‌اند (Rosario, 2000).

در ادامه، ابتدا مفاهیم پایهٔ مورد نیاز برای درک نمایه‌سازی معانی پنهان، و سپس مزایا و معایب نمایه‌سازی معانی پنهان را در کاربردهای مختلف بررسی می‌کنیم.

D_1, D_2=documents: بردار اسناد

q_1, q_2=queries: بردار درخواست

t_1, t_2=term



تصویر 1. مدل فضای بُرداری و زاویهٔ میان بُردار درخواست و سند

Doc_1	$term_{11}$	$term_{12}$	...	$term_{1i}$
Doc_2	$term_{21}$	$term_{22}$	...	$term_{2i}$
.	.	.	.	.
Doc_j	$termj_1$	$termj_2$		$term_{ji}$

تصویر 2. ماتریس اصطلاح ـ سند

[6] . Latent Semantic Indexing (LSI)

[7] . Singular Value Decomposition (SVD)

مفاهیم پایه

نمایه‌سازی معانی پنهان، روشی است که اسناد و درخواست را به فضایی با ابعاد معنایی پنهان منتقل می‌کند. در فضای معانی پنهان، سند و درخواست می‌توانند شباهت کسینوسی (کسینوس زاویهٔ میان دو بردار، که هرچه به 1 نزدیک‌تر باشد زاویهٔ میان دو بردار کمتر است) زیادی داشته باشند، حتی اگر از نظر لغوی، اصطلاحات مشترک نداشته باشند؛ درواقع شباهت میان معنای اصطلاحات، سنجیده می‌شود. می‌توان به روش نمایه‌سازی معانی پنهان، به عنوان یک استاندارد شباهت، برای رفع مشکل یکسان‌بودن لغوی اصطلاحات در فرمول TF*IDF توجه کرد (Rosario, 2000).

فضای معانی پنهان، که درخواست و سند را درونش پیاده‌سازی می‌کنیم، دارای ابعاد کمتری نسبت به فضای اولیهٔ اسناد است. ابعاد را می‌توان اصطلاحات کلیدی درنظر گرفت. در نتیجه می‌توان نمایه‌سازی معانی پنهان را روشی برای کاهش ابعاد (پیچیدگی) دانست. روش کاهش ابعاد، دارای یک مجموعه از اشیاء است که در یک فضا با بُعد بالاتر، موجودند و ما اعضای مجموعه را در فضایی با ابعاد پایین‌تر پیاده‌سازی می‌کنیم. برای درک بهتر می‌توان مثال فضای سه‌بُعدی و دوبُعدی را درنظر گرفت(Rosario, 2000).

نمایه‌سازی معانی پنهان، کاربردی از یک تابع محاسباتی ریاضی می‌باشد، که «تجزیهٔ مقدار ویژه برای ماتریس کلمه ـ سند» نامیده می‌شود. بنابراین، پایهٔ روش نمایه‌سازی معانی پنهان، محاسبهٔ کوچکترین مربعات از نظر ریاضی است. پیاده‌سازی در محیط معنایی پنهان، نسبت به فضای اولیه کوچکتر است، زیرا با درنظرگرفتن تفاضل کوچکترین مربعات، حداقل ابعاد برای پیاده‌سازی به‌دست آمده است. روش تجزیهٔ مقدار ویژه از ماتریس A، ماتریس $\bar{A}$ را محاسبه می‌کند و تفاضل دو ماتریس بوسیلهٔ نُرم[8] دوم، کمینه می‌گردد: (Ming, 1994)

$$\Delta=\|A-\bar{A}\|$$

نُرم دوم برای ماتریس، برابر با فاصلهٔ اقلیدسی میان بردارها است. تجزیهٔ مقدار ویژه، فضایی با N بُعد را به فضایی با K بُعد تبدیل می‌کند، و K از N بسیار کوچکتر است. در کاربرد بازیابی اطلاعات و ماتریس اصطلاح ـ سند، N تعداد اصطلاحات موجود در مجموعه است. مقدار K معمولاً بین 100 تا 150 انتخاب می‌گردد. نمایه‌سازی معانی پنهان، بردارها را از یک فضای N بُعدی به یک فضای K بُعدی منتقل می‌کند.

روش‌های گوناگونی برای انقیاد از فضایی با ابعاد بیشتر، به فضایی با ابعاد کمتر وجود دارد. نمایه‌سازی معانی‌پنهان روشی را برمی‌گزینند که مقدارکمینه برای Δ حاصل گردد. پس از محاسبهٔ«تجزیهٔ مقدارویژه ماتریس اصطلاح ـ سند»، A_{t*d} به سه ماتریس T_{t*n} و S_{n*n} و D_{d*n} تجزیه‌می‌گردد:(Rosario,2000;Ming,1994)

$$A_{t*d}=T_{n*n}S_{n*n}(D_{d*n})^T$$

t تعداد اصطلاحات، d تعداد اسناد و n مقدار کمینهٔ t و d است و ماتریس‌های T و D دارای ستون‌های نرمال[9] هستند. به عبارت دیگر، حاصل‌ضرب آن‌ها در ترانهاده[10]اش ماتریس واحد I، و رتبهٔ[11] ماتریس A برابر r می‌باشد.

انتخاب مقدار K برای $\bar{A}$ مسئلهٔ جالبی است. کم‌کردن مقدار K در برطرف‌کردن نویز هر بردار موثر است؛ ولی از طرف دیگر، ذخیرهٔ ابعاد پایین، اطلاعات مهمی را حذف می‌کند. به‌دست آوردن مقدار بهینه برای K در کاربرد، روش آزمون و خطا است. می‌توان ثابت کرد که تجزیهٔ مقدار ویژه برای یک ماتریس، منحصربه‌فرد است. بردارها نیز مانند اسناد، در فضای K بُعدی جدید پیاده‌سازی می‌شوند. برای انقیاد بردار ماتریس از فضای N بُعدی به فضای K بُعدی از این فرمول استفاده می‌شود: $q=Q^T T_{t*k} S^{-1}{}_{k*k}$

---

[8] . Norm
[9] . Ortho normal
[10] . Transpose
[11] . Rank

بهنگام‌سازی اطلاعات

برای به‌روز نگه‌داشتن اطلاعات در فضای جدید، باید روشی برای تزریق درخواست‌ها و اسناد تازه به فضای جدید انتخاب گردد. به‌دلیل پرهزینه‌بودن، مقرون‌به‌صرفه نیست که مجدداً برای ماتریس تغییریافته، تجزیهٔ مقدار ویژه محاسبه شود. بنابراین اصطلاحات و اسناد جدید به درون ماتریس موجود تزریق می‌گردند. برای تزریق اسناد در ماتریس می‌توان از فرمول زیر استفاده کرد:

$$A=TSD^T$$
$$T^T A=T^T TSD^T$$
$$T^T A=SD^T$$

بنابراین فقط کافی است که بردار درخواست یا سند، در ترانهادهٔ ماتریس اصطلاح T ضرب شود. پس از آن، ابعادش کاهش می‌یابد و به فضای دلخواه منتقل می‌گردد (Rosario, 2000; Hong, 2000).

مزایا و معایب

مزایا

در نمایه‌سازی معانی پنهان، فرض بر آن است که فضای جدید، محیطی مناسب‌تر از محیط اصلی برای پیاده‌سازی اسناد و درخواست‌ها می‌باشد. به‌کاربردن کلمهٔ «پنهان» در این روش استعاره از آن است که ابعاد جدید، پیاده‌سازی درستی می‌باشد. این پیاده‌سازی از آن نظر درست می‌باشد که در این ابعاد خاص، مجموعه‌ای از اصطلاحات در برخی اسناد و مجموعهٔ متمایز دیگری در اسناد دیگر، پیاده‌سازی را به عهده دارند. روش نمایه‌سازی معانی پنهان، ابعاد اولیه و ساختار معنایی آن‌را ترمیم می‌کند. بنابراین بزرگترین مزیت روش نمایه‌سازی معانی پنهان، در اصطلاح‌هایی است که مترادف، دارای چند معنا، و وابسته به اصطلاحات دیگر هستند (Rosario, 2000; Hong, 2000).

مترادف‌بودن، به این معنا است که یک مفهوم را می‌توان با اصطلاحات مختلف بیان کرد. انجام روش‌های کلاسیک بازیابی، با کلماتی که در اسناد مختلف، با معانی یکسان و تلفظ متفاوت به‌کار می‌روند. دشوار است. در روش نمایه‌سازی معانی پنهان، ارتباط معنایی میان این گونه اصطلاحات، در وزن جدیدی که شاخص محاسبه می‌نماید، درنظر گرفته می‌شود(Rosario, 2000; Hong, 2000).

تعدد معانی یعنی این که اصطلاحاتی هستند که بیش از یک معنا دارند و این، یک ویژگی عام در تمام زبان‌ها است. وجود اصطلاحاتی که تعدد معنا دارند، دقت نظام بازیابی را کاهش می‌دهد. روش نمایه‌سازی معانی پنهان در فضای جدید، نویز را کمینه می‌کند. بردار اصطلاحات نمایه‌سازی معانی پنهان، میانگین وزنی[12] معانی مختلف اصطلاح می‌باشد. وقتی معنای واقعی با میانگین معانی متفاوت باشد، نمایه‌سازی معانی پنهان در عمل، کیفیت جستجو را کم می‌کند (Rosario, 2000; Hong, 2000).

مدل برداری کلاسیک فرض می‌کند که اصطلاحات مستقل، به صورت بردارهای پایهٔ عمود برهم هستند. به دلیل پیوستگی میان اصطلاحات در زبان، این فرض هیچ‌گاه صحیح نیست. استقلال یا پیوستگی اصطلاحات، اگر به درستی مورد توجه قرار گیرد، کارآیی جستجو را افزایش می‌دهد. اضافه‌کردن عبارت[13] به اجزای جستجو، از کاربردهای این قسمت است.

معایب

به نظر می‌رسد که تجزیهٔ مقدار ویژهٔ ماتریس، دارای حجم کمتری نسبت به مدل اولیه باشد. مثلاً اگر عدد 150 را برای پارامتر K درنظر بگیریم، طول بردارها 150 می‌باشد، ولی در عمل مشکل دیگری وجود

---

[12] . Weighted Average
[13] . Phrase

دارد. در حالتی که بردارها به روش معمول در فضای اولیه پیاده‌سازی می‌گردند، ماتریس و بردارها، به صورت ماتریس پراکنده[14] ذخیره می‌شوند، ولی پس از انتقال بـه ابعـاد جدیـد، دیگـر حالـت پراکنـده ندارنـد. مـشکل ذخیره‌سازی بر روی دیسک، امروزه قابل حل است، ولی برخی از عملیات ریاضی، روی ماتریس‌های پراکنـده سریعتر انجام می‌گیرند (Rosario, 2000; Hong, 2000).

یکی از مشکلات مهم در به‌کارگیری نمایه‌سازی معانی پنهان، کارآیی سامانه از نظر سرعت پاسخگویی است. سامانه‌هایی که از فضای برداری در حالت کلاسیک استفاده می‌کنند، نمایهٔ مقلـوب را در زمـان بازیـابی برای مقایسهٔ بردارها به‌کار می‌برند. بنابراین فقط اسنادی که اصطلاحات کلیدی درخواست را شامل مـی‌شـوند، مقایسه می‌گردند. ولی در روش نمایه‌سازی معانی پنهان، بردار درخواسـت بایـد بـا تمـام بردارهـای موجـود در مجموعه مقایسه گردد. بنابراین ضرب بردارها برای محاسبهٔ شباهت و زاویهٔ میان بردارهـا، زمـان بیـشتری طلـب می‌کند (Rosario, 2000; Hong, 2000).

تصویر 3 مقایسه‌ای میان روش نمایه‌سازی پنهان و دیگر روش‌های کلاسیک را نشان می‌دهـد. مـشاهده می‌گردد که این روش، کارآیی بیشتری از مدل‌های دیگر دارد. این آزمایش به روی مجموعهٔ «گرانفیلـد» کـه دارای 1400 سند و 225 درخواست می‌باشد انجام شده است.

کاربردهای نمایه‌سازی معانی پنهان

بازیابی اطلاعات: کاربرد نمایه‌سازی معانی پنهان در بازیابی اطلاعات، و در به‌دست‌آوردن لیست اسـناد مرتبط با درخواست است. کارآیی این روش در مجموعه‌های مختلف آزمایش شده است و نتایج به‌دست آمده نشان می‌دهد که در روش نمایه‌سازی معانی پنهان، میانگین دقت حدود 30 درصد بهتر اسـت (Foltz, 1998; Dumais, 1997).

ارتبـاط بـازخورد: در ایـن روش، بـرای اعمـال بـازخورد از اسـناد انتخـاب‌شـده توسـط کـاربر، بـردار درخواست با مجموع بردارهـای درخواسـت‌هـای انتخـاب‌شـده بـه عنـوان مـرتبط در لیـست پاسـخ، جـایگزین می‌گردد. اگر بردار درخواست، با اولین سند شبیه، جایگزین گردد، کارآیی سیـستم حـدود 33 درصـد اضـافه می‌شود، و جایگزینی بردار درخواست با میانگین اولین سه سند شبیه، حدود 67 درصد کارآیی کل سیـستم را اضافه می‌کند. پس این روش به عنوان کاربرد نمایه‌سازی معانی پنهان در بهینه‌سازی درخواسـت، مـورد توجـه است (Foltz, 1998; Dumais, 1997).

پالایش کردن اطلاعات: کاربرد نمایه‌سازی معانی پنهان در جداسازی اطلاعـات، بـسیار سـاده اسـت. در این کاربرد، بردار سند انتخابی توسط کاربر با بردار اسناد دیگر مقایسه می‌گردد و معیار شـباهت اسـناد بـا یـک مقدار آستانه،[15] اسناد شبیه را از میان اسناد پالایش خواهد کرد. در این میان به‌کارگیری روش‌هـای هوشـمند و بازخورد، نتایج خوبی را حاصل خواهد کرد (Foltz, 1998; Rosario, 2000; Dumais, 1997).

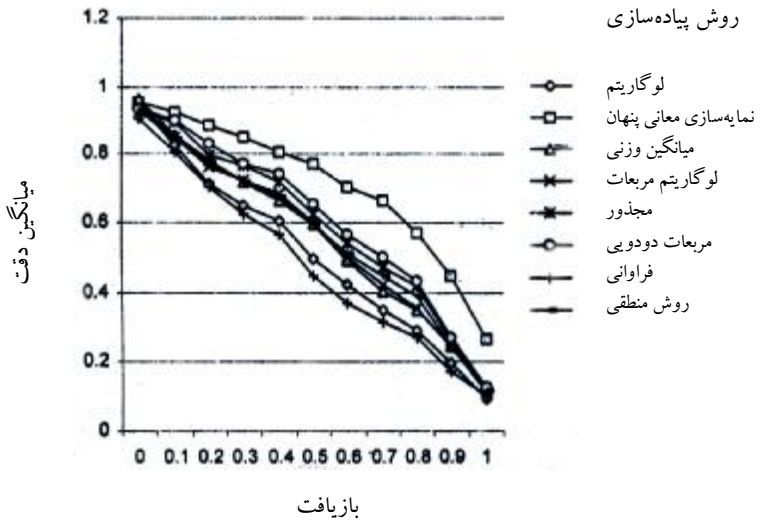
بازیابی مستقل از زبان: نکتهٔ مهم در به‌کارگیری نمایه‌سازی معانی پنهان آن است که این روش از قواعد لغوی یا معنایی مربوط به زبان انگلیسی استفاده نمی‌کند، بنابراین برای بازیابی اطلاعات در هر زبان مـی‌تـوان از این روش به منظور بازیابی اسناد مشابه استفاده کرد. امکانات مورد نیاز فقط فضای مـشترک بـرای پیـاده‌سـازی اصطلاحات و بردارها است، بنابراین می‌توان اسناد و ماتریس اصطلاح- سند را بـرای مـدارکی کـه در بـیش از یک زبان دارای نسخهٔ متنی هستند مورد استفاده قرار داد. مثلاً می‌توان از زبان انگلیسی و فرانسوی استفاده کرد

---

[14] . Sparse Matrix
[15] . Threshold

و یک فضای مشترک برای هر دو زبان ایجاد کرد، و در این حالت نیاز به ترجمهٔ درخواست نیست و مشکلی در بازیابی اسناد مرتبط به وجود نخواهد آمد(Foltz, 1998; Rosario, 2000; Dumais, 1997).

ورودی‌های نویزی: به‌دلیل آن‌که روش نمایه‌سازی معانی پنهان به تلفظ اصطلاحات کلیدی وابسته نیست، در مواردی که همراه با نویز باشد، بسیار مفید است. مثلاً زمانی که املای اصطلاح غلط باشد یا در شناخت نویسه‌های تصویری،[16] می‌توان این روش را به‌کار برد. اگر غلط املایی فقط در یک مکان وجود داشته باشد در فضای جدید اسناد، ترمیم می‌گردد و با نسخهٔ صحیح از اصطلاح، در میانگین مربعات جایگزین می‌گردد (Foltz, 1998; Rosario, 2000; Dumais, 1997).



تصویر 3. مقایسهٔ کارآیی میان روش‌های پیاده‌سازی بازیابی اطلاعات در مدل برداری

نتیجه

نمایه‌سازی معانی پنهان روشی تازه و امیدبخش در بازیابی اطلاعات است، که اسناد را در فضایی با ابعاد کمتر، شاخص‌گذاری و بازیابی می‌کند. هدف از تغییر فضای اسناد، استفاده از نمایه‌سازی و بازیابی معنایی اصطلاحات است. در این روش برای کمینه‌کردن نُرم ماتریس، از تجزیهٔ مقدار ویژه استفاده می‌شود و در مقایسه‌های انجام‌شده در TREC[17] جایگاه مناسبی به‌دست آورده است. این روش به‌خصوص زمانی که نمایه‌سازی و جستجوی مستقل از زبان، در اولویت باشد، کارآیی مناسبی خواهد داشت. مثلاً موتور جستجوگر «گوگل» از نمایه‌سازی معانی پنهان، در نمایه‌سازی‌ها و جستجوهای مستقل از زبان استفاده می‌کند.

منابع

Kowalski, G. "Information retrieval systems, Theory and Implementation", Kluwer Publisher. 1997. [Online] Available: http://gunther.smeal.psu.edu/context/48183/0

Salton G., McGill M J. "An introduction to Information Retrieval", McGraw Hill, 1983 [Online] Available: http://citeseer.nj.nec.com/context/1241/0

Berry M. W., Dumais S T., Shippy A. T., "A case study of latent semantic

[16] . OCR (Optical Character Recognition)

indexing", Tech. Rep. CS-95-271, University of Tennessee, Knoxville, January 1995.

ROSARIO B., "Latent Semantic Indexing: An overview", INFOSYS 240, spring 2000.

Ming Gu, Demmely James, Dhillonz Inderji, "Efficient Computation of the Singular Value Decomposition with Applications to Least Squares Problems", Lawrence Berkeley National Laboratory Technical Report LBL-36201, September 29, 1994 [Online] Available: www.cs.berkeley.edu/~inderjit/fac_resume.ps

Hong Jason I., "An Overview of Latent Semantic Indexing", SIMS 240, spring 2000.

PETER W. Foltz, "Using Latent Semantic Indexing for Information Filtering" IEEF-CS\TCDA, University of Colorado, 1998.

Dumais Susan, Bellcore, "Using Latent Semantic Indexing (LSI) For Information Retrieval, information Filtering and Other Things", Cognitive Technology Conference Cognitive Technology Conference April 4, 1997.