سال دوم، شماره ۱ (پاییز ۱۳۸۳)

اطلاع شناسی
INFORMOLOGY

چکیده

در این مقاله تکنیک‌های بازخورد ربط با استفاده از روش مدل‌سازی رتبه‌ای و به شکلی ساده مطرح می‌شود و نمودارهای مربوط، مفهوم روش مدل‌سازی زبان در بازیابی اطلاعات را شفاف‌تر می‌سازد.

سه مسئله‌ای که انتظار می‌رود یک مدل بازیابی اطلاعات بدان پاسخ دهد عبارت‌اند از: ۱) وزن‌دهی به لغات؛ ۲) بازخورد ربط؛ ۳) پرسش‌های ساختارمند. بر اساس این پرسش‌هاست که رخداد ربط میان منابع متفاوت سنجیده می‌شود.

در متون، از مدل‌های بازیابی قابل رتبه‌بندی متعددی نامبرده شده است که مقالهٔ حاضر به توصیف اجمالی برخی از آنها و ذکر مزایا و معایب آنها می‌پردازد.

کلیدواژه‌ها: ربط. بازخورد ربط. مدل‌سازی رتبه‌ای. بازیابی اطلاعات.

بررسی اجمالی مزایا و معایب مدل‌های بازخورد ربط

فاطمه فهیم‌نیا

بررسی اجمالی مزایا و معایب مدل‌های بازخورد ربط

فاطمه فهیم‌نیا[1]

آنچه در مقابل کلمهٔ مدل در واژه‌نامهٔ وبستر آمده عبارت است از: "نظامی مفروض برای توصیف ریاضی داده‌ها، نتایج، ماهیت یا وضع امور". تعبیر دومی نیز به این صورت آمده است: "مدل، الگویی است برای چیزی که می‌خواهد ساخته شود" (وبستر[2]، ۱۹۸۳).

مدل بازیابی اطلاعات، آنچه را که کاربر مرتبط با پرسش خود از نظام دریافت خواهد کرد شرح می‌دهد و پیش‌بینی می‌کند. از این رو ضروری است که صحت عملکرد مدل‌های بازیابی اطلاعات در محیطی کنترل شده مورد آزمایش قرار گیرد.

بسیاری از مدل‌های بازیابی اطلاعات، که در اینجا مورد بحث قرار می‌گیرند، ماهیت الگو بودن را برای ساخت نظام‌های جدید دارند و ویژگی‌های آنها باعث گزینش از سوی یک نظام برای بازیابی یا سنجش عملکرد نظام‌های ذخیره و بازیابی اطلاعات خواهد بود. این مدل‌ها به منزلهٔ راهنمایی برای ایجاد این‌گونه نظام‌ها به‌کار گرفته می‌شود.

شاید در مجموعه‌های کوچک، انجام پویشی خطی کفایت کند، لیکن هنگامی که در ورای متن، داده‌ها ساختار دیگری یابند نظیر ایجاد فایل‌های مقلوب یا فایل واژگانی ـ

۱. دانشجوی دکترای کتابداری و اطلاع‌رسانی دانشگاه تهران
Fahimnia@ut.ac.ir

2. Webster

در زمانی که مجموعه مدارک وسیع‌تر است ـ این شیوه کارآیی چندانی ندارد.

از آنجایی که تولید فایل‌های مقلوب یا واژگانی، مانند نمایهٔ آخر کتاب، برای نظام بازیابی اطلاعات الزامی است، مدل‌سازی بازیابی اطلاعات نیز با همان ضرورت صورت می‌پذیرد. فایل مقلوب معمولاً شناسه‌های متن را گاه با محل قرارگیری آن در متن و گاه حتی با ذکر وزن آن در متن ذکر می‌کند.

معمولاً ساختار داده‌ها از دو زیرساخت ضروری نشأت می‌گیرد (هیمسترا[1]، ۲۰۰۲):

۱. فایل واژگانی که شامل لغات متن است.

۲. فایل نشانی که شامل رخداد لغات در متن است.

سه مسئله‌ای که انتظار می‌رود هر مدل بازیابی اطلاعات بدان پاسخ دهد عبارتند از:

۱. وزن‌دهی به لغات. به عبارت دیگر، مدل‌هایی که، یک الگوریتم وزن‌دهی به کلمات را حائز اهمیت می‌شمارند مناسب‌ترین‌اند. وزن یک لغت میزان اهمیت آن کلمه در متن را نشان می‌دهد و ساده‌انگاری در مورد آن، مدل را از قوت می‌اندازد.

۲. بازخورد ربط که براساس مثال‌هایی در متون مرتبط سنجیده می‌شود و این اساس میزان اعتبار مدارک و متون دیگر را تخمین می‌زند.

۳. پرسش‌های ساختارمند، به نحوی که یک پرسش آمیزه‌ای از کلمات انگاشته نشود. در برخی مدل‌ها بر اساس این پرسش‌هاست که رخداد ربط میان منابع متفاوت نیز سنجیده می‌شود.

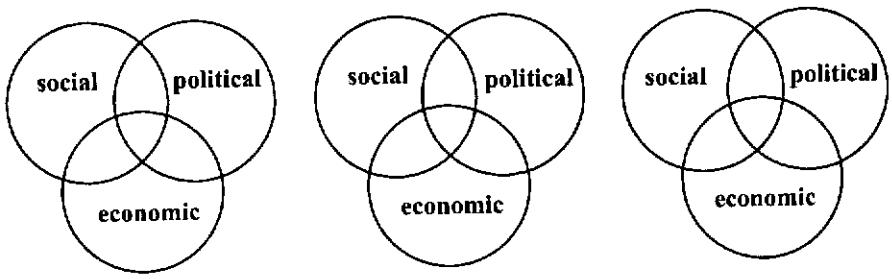
در متون، از مدل‌های بازیابی متعددی نامبرده شده است که تنها ۸ مدلی که در ذیل به توصیف اجمالی آنها فارغ از مباحث ریاضی می‌پردازیم، مسائل مذکور را در ساخت مدل به طریقی لحاظ کرده‌اند.

مدل‌های بازیابی فاقد قابلیت رتبه‌بندی نتایج

۱. مدل بولی[2]

این مدل که بر اساس عملگرهای منطقی or ، and و not شکل گرفته است از اولین مدل‌های بازیابی اطلاعات است که بیش از هر مدل دیگری مورد استفاده یا نقد قرار گرفته است. اساس کار این مدل در نمودارهای ون[3] مندرج در شکل زیر آمده است (چودری[4]، ۱۹۹۸).

1. Hiemstra
2. Boolean
3. Venn Diagrams
4. Chowdhury



مزایای این مدل را می‌توان به شکل زیر ذکر کرد (راسموسن[1]، ۱۹۹۹).

۱. ایـن مـدل بـه کـاربران خبره‌ای کـه بـا آن کـار مـی‌کنند، احسـاس تسـلط بـه سیستم را می‌دهد. وقتی که تعداد مدارک بـازیابی شـده بسـیار انـدک یـا خـیلی زیـاد بـاشد بـه راحـتی بـا اسـتفاده از عـملگرهای فـوق مـی‌توان پـرسش را ویـرایش کرد.

۲. اکنون این مدل توسط عملگرهای نزدیک‌یابی[2] و کوتاه‌سازی[3] به روشی منطقی و ریاضی برای کاوش متن کامل نیز تقویت شده است.

۳. از همه مهم‌تر، کاربران با این عملگرها در طی دورانـی کـه کـاوش در نـظام‌های اطلاعاتی باب شده است آشنا شده‌اند. کاربران در طی زمان، استفاده از آن عـملگرها را بیش از مدل‌های دیگر فراگرفته‌اند.

۴. هزینه‌های اصلی در طراحی و ساخت نرم‌افزار متناسب با این عـملگرها بسـیار ناچیز است.

این مدل مانند همهٔ مدل‌ها معایبی نیز دارد (ساوینو[4] و سباستیانی[5]، ۱۹۹۸):

۱. برای کاربر مبتدی مشکلات مدل‌های دیگر را دارد؛

۲. قادر به رتبه‌بندی نتایج بازیابی شده نیست؛

۳. تفاوت ریاضی عملگرهای بولی در زبان طبیعی با این شکل لحاظ نمی‌شود. در واقع، این مدل بسیار پیچیده‌تر از نیازهای کاربر مبتدی است.

پس، این مدل فاقد توان رتبه‌بندی مدارک بـازیابی شـده است و در حـال حـاضر بـا نـیازهای بـازیابی مـتن کـامل مـنطبق نـیست مگـر بـه پشـتوانـه عـملگرهای افزوده.

بسیاری بر این باورند که این مدل «مدل مدل‌ها» است؛ یعنی مناسب آن است کـه کاربر با آن کار کند ولی مدل‌های دیگر در زیرساخت سیستم با رتبه‌بندی مدارک به‌کار

1. Rasmussen
2. Proximity
3. Wild cards
4. Savino
5. Sebastiani

گرفته شوند.

مدل‌هایی که از این پس به آنها اشاره می‌شود مدل‌های بازیابی دارای قابلیت رتبه‌بندی هستند.

مدل‌های بازیابی با قابلیت رتبه‌بندی نتایج

این مدل‌ها معمولاً از روش‌های آماری برای سنجش بسامد کلمه در متون استفاده کرده و بر آن اساس رتبه را محاسبه می‌نمایند. اغلب این مدل‌ها از پرسش‌های ساختارمند بهره می‌گیرند:

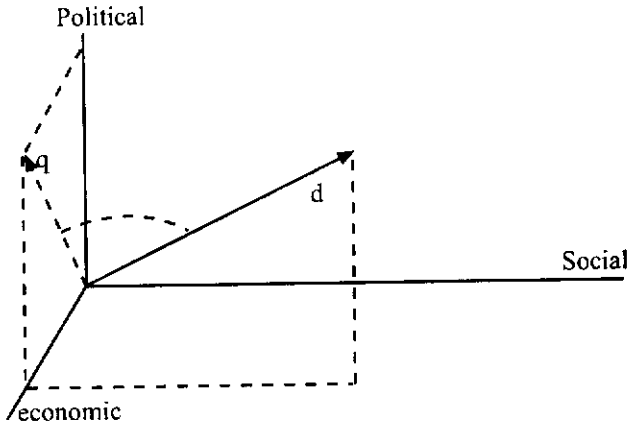
۱. مدل بُرداری[1]

بر اساس این مدل زمانی که پرسش، حاوی کلمه‌ای باشد که در متن بایستی بازیابی شود، از آنجایی که هر کلمه در پرسش امکان رخداد مخصوص خود را دارد که ممکن است با امکان رخداد آن در متن متفاوت باشد، کلمات موجود در پرسش به عنوان یک بردار تلقی شده و بسامد آن در متن بردار دوم را تشکیل می‌دهد. سنجش برآیند این دو بردار، متون مرتبط را شناسایی می‌کند. این مدل در سال ۱۹۵۷ توسط لوهن[2] پیشنهاد شد و از اولین مدل‌های بازیابی محسوب می‌شود؛ اما به دلیل فقدان کاربردهای جاری ریاضی و ضعف مبانی نظری و نیز به دلیل عدم ارائهٔ مقیاس برای هر بردار، کاربرد زیادی نیافت (ون ریسبرگن[3]، ۱۹۷۹).

۲. مدل بُرداری فضایی[4]

سالتون و مک‌گیل (۱۹۸۳) بر اساس مدل لوهن این مدل را که از زیربنای منطقی قوی‌تری بهره‌مند است ارائه کردند. در این مدل فرض بر آن است که بردارهای حاصل از پرسش و متن در فضایی اقلیدسی موجودند؛ از این رو، هر کلمه مختصات خود را در این فضا خواهد داشت که معمولاً مقدار آن برابر کسینوس زاویه‌ای است که بین دو بردار پرسش و متن ایجاد شده است (ویتن[5] و همکاران، ۱۹۹۴) بر اساس این ساختار فضایی، تصور مدل بسیار آسان است و نمونهٔ آن در زیر آمده است:

1. Vector
2. Luhn
3. Van Rijsbergen
4. Vector space
5. Witten



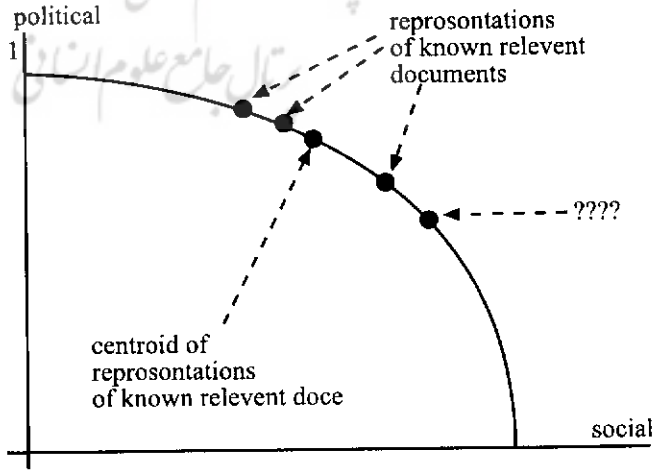
معایب این مدل عبارتند از:

۱. مدل بردار فضایی نشان‌دهندهٔ ارزش هر بردار و اجزای آن نیست (سالتون[1]، ۱۹۷۱)؛

۲. محاسبهٔ کسینوس برای تمامی بردارها عملی نیست، چون مقدار دقیق بردارها لازم است؛

۳. ارتباط و وابستگی لغات در این روش نشان داده نمی‌شود.

به‌منظور رفع نقص اول و دوم اقدام به نرمالیزه کردن روش، توسط تخمین بردارها، پایه و اساس روش روکیو است که در عمل مبتنی بر مقدار لگاریتم هر بردار است. نمودار مثالی از کاربرد این مدل نیز در ذیل آمده است (ساوینو و سباستینی، ۱۹۹۸).



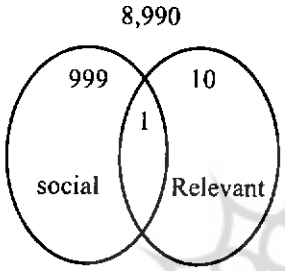
1. Salton

۳. مدل احتمالی[1]

مارون[2] و کنز[3] در سال ۱۹۶۰، مدلی را ارائه کردند که هدف آن رتبه‌بندی مدارک موجود در مجموعه و میزان ربط آنها با پرسش بر اساس قوانین احتمالات بود.

در اینجا سؤال اصلی آن است که چگونه و بر اساس کدام داده‌ها، احتمال ربط بایستی تخمین زده شود.

رابرتسون[4] در ۱۹۷۷، مدل بولی را از طریق نمودار زیر با مدل احتمال ربط تلفیق کرده است:



فرض بر این است که لغات جدای از هم و مدارک نیز به‌صورت مستقل سنجش احتمال می‌شوند (رابرتسون، ۱۹۷۶؛ و ون ریسبرگن، ۱۹۷۹). مدل احتمال‌سنجی یکی از معدود مدل‌های بازیابی است که نیاز به الگوریتم وزن کلمات افزوده به پرسش ندارد؛ و از این رو، یکی از روش‌های بسیار مؤثر در بازیابی اطلاعات است (کرافت و هارپر، ۱۹۷۹).

متأسفانه در حال حاضر نرم‌افزارهای کاربردی متعددی برای به‌کارگیری این مدل وجود ندارد و بزرگ‌ترین نقص آن این است که رتبه‌بندی دقیق از متن به دست نمی‌دهد (فوهر[5]، ۱۹۹۲).

۴. مدل لمّایی (فازی)[6]

در مدل لمّایی، که در سال ۱۹۶۵ توسط زاده[7] ارائه شد، هر مدرک درجه‌ای از عضویت در یک مجموعه دارد. در این مدل، مدرک درجه‌بندی شده در هر مجموعه با کلمه‌ای در نمایه شناسایی می‌شود.

مزیت این مدل، در قیاس با مدل احتمال‌سنجی و بردار فضایی، در فراهم‌آوردن

1. Probabilistic
2. Maron
3. Kohns
4. Robertson
5. Fuhr
6. Fuzzy
7. Zadeh

امکـان رتـبه‌بندی پـرسش‌های سـاختارمند است (لی[1]، ۱۹۹۵)؛ لیکـن مـانند مـدل برداری برای وزن‌دهی به کلمات برای هر کلمهٔ اضافی در پرسش به الگوریتم خـاصی نیاز است. مـنطق لمّـایی بـه دلیـل مـاهیتش از تـوجیه ایـنکه چـرا یک عـملگر بـهتر از عملگرهای دیگر عمل می‌کند و اینکه در پشت این مدل واقعاً چه می‌گذرد نـاتوان است.

۵. مدل بسط یافتهٔ بولی p-norm

این مدل که در سال ۱۹۸۳ توسط سالتون، فاکس[2]، و وو[3] ارائه شد، بر پایهٔ مدل برداری فضایی اقلیدسی بنا شده و در آن استفاده از یک ضـریب نـرمال‌سازی ρ بـرای سنجش وزن کلمات پیشنهاد شد که فرمول‌های سنجش برداری را بهینه ساخت. در حال حاضر، این مدل یکی از عمومی‌ترین مدل‌های کاربردی است که در معماری بسیاری از نظام‌های بازیابی به‌کار گرفته شده است (گریف[4]، کرافت[5]، و تورتل[6] ۱۹۹۷). همچنین مانند مدل لمّایی برای هر کلمهٔ اضافی در پرسش نیاز به الگوریتم خاص دارد (لوسادا[7] و باریرو[8]، ۱۹۹۹).

۶. مدل 2-Poisson

در سال ۱۹۷۴ به‌منظور مطالعهٔ قواعد آماری لازم برای شناسایی کلمات نمایه در یک متن، بوکشتاین[9] و سوانسون[10] پیشنهاد کرده‌اند که اگر تعداد رخداد کلمات در متن با تـلفیقی از تـوزیع 2-Poisson مـدل‌سازی شـود، بـرای هـر کـلمه، مـجموعه بـه دو زیرمجموعه تقسیم مـی‌شود. مـدارک در زیـرمجموعهٔ اول بـه مـوضوعی ارجـاع داده می‌شوند که تـوسط کـلمات مـوجود در زیـرمجموعهٔ دوم ــ کـه مـنطقاً کـوچک‌تر از مجموعه اول است ــ قابل بازیابی است.

مزیت اصلی این مـدل آن است کـه نـیاز بـه الگـوریتم وزن‌دهـی بـرای هـر کـلمهٔ افزون ندارد. لیکن شکل اصلی در آن تخمین پارامـترهاست کـه در مـدل بـه‌کـارگیری می‌شود. برای هر کلمه سه پـارامـتر مـوجود است کـه بـه‌طور مسـتقیم از روی تـعداد رخدادهای کـلمات قـابل تـخمین نـیست. عـلی‌رغم پـیچیدگی ایـن مـدل، هـنوز هـم مـتونی کـه دارای بسـامد کـلمات فـراوان هسـتند بـا ایـن مـدل قـابل تـخمین نیستند.

1. Lee
2. Fox
3. Wu
4. Greiff
5. Croft
6. Turtle
7. Losada
8. Barreiro
9. Bookstein
10. Swanson

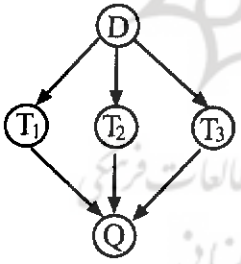
۷. بسط مدل احتمالی

رابـــرتسون، ون ریســـبرگن، و پـــورتر[1] (۱۹۸۱) مـــدل 2-Poisson را بـــا مـــدل احتمال‌سنجی تلفیق کردند. به این منظور الگوریتم وزن‌دهی به کلمات افزون را با مدل قبلی ترکیب کرده و ضرایب مربوط را تـغییر دادنـد. بـرای نـمایهٔ مـناسبی از کـلمات، می‌بایست مدارک مرتبط و نامرتبط به‌خوبی شناسایی و انتخاب گردند. احـتمال ایـن انتخاب نیز از طریق این مدل قابل سنجش است. از این رو، میزان ربط سنجیده شده از این طریق بر سنجش بسامد کلمات برتری دارد.

مزیت اصلی این مدل نیز عدم نیاز به الگوریتم وزن‌دهی برای کلمات اضافی است (هارتر[2]، ۱۹۷۵) لیکن در کاربرد ضرایب در محاسبه به طریقی عمل شده که گویی طول مدارک یکسان است؛ و این امری نادر است که مدارک از نظر تعداد کلمات موجود در متن یکسان باشند.

۸. مدل‌های مبتنی بر شبکه Bayesian

شبکهٔ **Bayesian** شبکهٔ بسته‌ای است که ارتباط وابسته به احتمال وقوع متغیرها را نشان می‌دهد. مدل ساده آن به شکل زیر است:



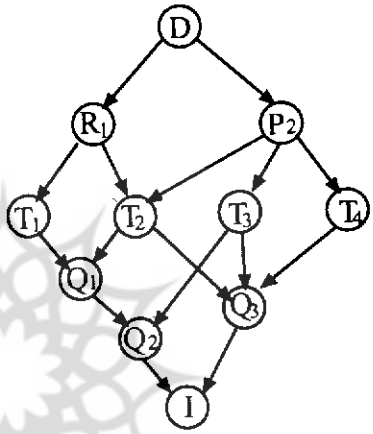
این مدل ساده نمایانگر آن است که، به‌طور مثال، یک مدرک از طریق ۳ کلمه بـا یک پرسش مرتبط می‌شود.

این مدل با به‌کارگیری محاسبات ریاضی، امکان بررسی ارتباط کلمات، متن، و پرسش را به نحوی مستقل از وابستگی کلمات به متن اصلی فراهم می‌سازد و با افزودن هر کـلمه در پرسش، مدل پیچیده‌تر می‌شود؛ چون احتمال آن بر احتمال کلمات قبلی افزوده می‌گردد.

مزیت اصلی این مدل، بر اساس آنچه تورتل و کرافت در ۱۹۹۲ ابراز داشته‌اند، آن

1. Porter
2. Harter

است که ریخت‌شناسی شبکه، رخداد کلمات را به طریقی تلفیقی با هم مرتبط می‌سازد. لیکن این مدل نیز معایبی دارد (ریبیرو[1] و مانتز[2]، ۱۹۹۶). محاسبهٔ احتمالات برای کلمات مختلف سخت و وقت‌گیر است (تورتل و کرافت، ۱۹۹۲) و این خود به شدت بر کارکرد این مدل در نظام‌های بازیابی اطلاعات تأثیر می‌گذارد (ون ریسبرگن، ۱۹۸۶؛ سباستینی، ۱۹۹۴؛ فوهر، ۱۹۹۵؛ و ونگ[3]، ۱۹۹۵). به‌طور مثال، افزایش تنها یک سطح به شبکه، آن را به نحو قابل توجهی پیچیده‌تر می‌سازد:



مآخذ

Bookstein, A. and D. R. Swanson (1974). Probabilistic models for automatic indexing. *Journal of the American Society for Information Science*, 25 (5):313-318.

Chowdhury, G. G. (1998). *Introduction to modern information retrieval*. John Wiley & Sons.

Croft, W. B. and D. J. Harper (1979). "Using probabilistic models of document retrieval without relevance information". *Journal of Documentation*, 35 (4): 285- 295.

Fuhr, N. (1992). Probabilistic models in information retrieval. *The Computer Journal*, 35 (3): 243-255.

Greiff, W. R., W. B. Croft, and H. R. Turtle (1997). Computationally tractable

1. Ribeiro
2. Muntz
3. Wong

probabilistic modeling of boolean operators. In Proceedings of the 20th ACM Conference on Research and Development in Information Retrieval (SIGIR, 97): 119-128.

Harter, S. P. (1975). "An algorithm for probabilistic indexing". *Journal of the American Society for Information Science*, 26 (4): 280-289.

Hiemstra, D.(2002). Using Language Models for Information Retrieval, Ph.D thesis- Enschede: Neslia Paniculata, Netherland.

Lee, J. H. (1995). Analyzing the effectiveness of extended boolean models in information retrieval. *Technical Report TR95-1501, Comel*1 University. htt//:llcs-tr.cs.comell.edul

Losada, D. E. and A. Barreiro (1999). Using a belief revision operator for document ranking in extended boolean models. In *Proceedings of the 22nd ACM Conference on Research and Development in Information Retrieval* (SIGIR, 99), pp.66-73.

Luhn, H. P. (1957). "A statistical approach to mechanised encoding and searching ofliterary information". *IBM Journal of Research and Development*, 1 (4): 309-317.

Maron, M. E. and J. L. Kuhns (1960). "On relevance, probabilistic indexing and information retrieval". *Journal of the Association for Computing Machinery*, 7, pp. 216-244.

Rasmussen, E. M. (1999). Libraries and bibliographical systems. In R. A. Baeza-Yates and B. Ribeiro-Neto (Eds.). Modern Information Retrieval, pp. 397-413. Addison-Wesley. Ribeiro, B. A. N. and R. Muntz (1996). A belief network model for IR. In *Proceedings of the 19th ACM Conference on Research and Development in Information Retrieval* (SIGIR, 96): 252-260.

Robertson, S. E. (1977). "The probability ranking principle in IR". *Journal of, Documentation*, 33 (4): 294-304.

Robertson, S. E. and K. Sparck-Jones (1976). "Relevance weighting of search

terms". *Journal of the American Society for Information Science*, 27, pp. 129-146.

Robertson, S. E., Van Rijsbergen, and M. F. Porter (1981). Probabilistic models of indexing and searching. in R. N. Oddy et al. (Eds.). *Information Retrieval Research*, pp. 35-56. Butterworths.

Salton, G. (1971). *The SMART retrieval system: Experiments, in automatic document processing*. Prentice-Hall.

Salton, G., E. A. Fox, and H. Wu (1983). "Extended boolean information retrieval". *Communications of the A.CM*, 26 (11): 1022-1036.

Salton, G. and M. J. McGill (1983). *Introduction to Modern Information Retrieval*. McGraw-Hi11.

Savino, P. and F. Sebastiani (1998). "Essential bibliography on multimedia information retrieval, categorisation and filtering". in *Slides of the 2nd European Digital Libraries Conference Tutorial on Multimedia Information Retrieval*.

Sebastiani, F. (1994). "A probabilistic terminological logic for modelling information retrieval". in *Proceedings of the 17th ACM Conference on Research and Development in Information Retrieval (SIGIR, 94)*: 122-130.

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

Van Rijsbergen, C. J. (1979). *Information Retrieval*, second edition. Butterworths. [online] Available: http//:1lwww .dcs.gla.ac. uk/Keith/ Preface. html

Van Rijsbergen, C. J. (1986). "A non-classical logic for information retrieval". *The Computer Journal*, 29 (6): 481-485.

Webster's Ninth New Collegiate Dictionary. Merriam-Webster Inc.

Witten, I. H., A. Mo_at, and T. C. Bell (1994). *Managing Gigabytes: Compressing and Indexing Documents and Images*. Van Nostrand Reinhold.

Zadeh, L. A. (1965). "Fuzzy sets". *Information and Control* 8, PP. 338-353.