بررسی مدل فضای برداری در بازیابی اطّلاعات

دکتر جعفر مهراد[۱]

مهندس سارا کلینی[۲]

چکیده

بازیابی اطّلاعات، فرایند یافتن اطّلاعات (مدارک) مربوط به جستجوی کاربر در مجموعهٔ مدارک است. با پیاده سازی الگوریتمهای متفاوت، استراتژیهای مختلفی در بازیابی اطّلاعات وجود دارد. وجه مشترک استراتژیهای بازیابی، یافتن مدارک مشابه با موضوع جستجوی کاربر است. یکی از الگوریتمهای مهمی که در بازیابی اطّلاعات، کاربرد بسیار دارد، الگوریتم فضای برداری است که می کوشد تمام مدارک را در مجموعه و جستجوهای کاربر به صورت بردارها نشان دهد و ضریب تشابه میان بردارهای مدرک و بردار جستجو را جهت بازیابی مدرک مربوط، محاسبه نماید.

کلیدواژه‌ها: بازیابی اطّلاعات، مدل فضای برداری، فراوانی اصطلاح، وزن اصطلاح، رتبه‌بندی مدارک.

مقدمه

پیشرفتهای اخیر در علم الکترونیک، به تولید ابزارهای مدرن برای ذخیره‌سازی انبوهی از اطّلاعات منجر گردیده است. انفجار اطّلاعات باعث شده است تا جامعهٔ پژوهشگران در حوزهٔ بازیابی اطّلاعات، امکان و شیوهٔ فراخوانی اطّلاعات درخواستی را در پنجاه سال اخیر بسیار بهبود ببخشند. [۲ و ۱] با سیستمهای بازیابی اطّلاعات امروزی، امکان جستجو در چند ترابایت اطّلاعات، فقط در چند ثانیه وجود دارد [۳].

---

۱. استاد بخش علوم کتابداری و اطلاع رسانی دانشگاه شیراز و رئیس کتابخانه منطقه ای علوم و تکنولوژی

۲. کارشناس ارشد مهندسی کامپیوتر و رئیس اداره فناوریهای اطّلاعاتی کتابخانه منطقه ای علوم و تکنولوژی

نظام بازیابی اطّلاعات به سازماندهی، ذخیره سازی، بازیابی و نمایش اطّلاعات کتابشناختی مربوط است. سیستمهای بازیابی اطّلاعات با هدف فراهم آوردن زمینهٔ لازم برای پاسخگویی به جستجوهای کاربر از طریق ارجاع به مدارک مربوط، طراحی می‌گردد. در چنین محیطی، مجموعه‌ای از مدارک مانند کتابها، مقاله‌ها، گزارشهای تحقیقاتی و... وجود دارد، به اضافهٔ گروهی از کاربران. نیاز اطّلاعاتی کاربر در یک زمان خاص می‌تواند شامل یک یا چند مدرک باشد. مفهوم «ربط»، عامل مورد توجّهی در مسئلهٔ بازیابی است.

یک مدرک با توجه به ویژگیهایی که دارد (نحوهٔ نگارش، موضوع و...) و یا با در نظر گرفتن مشخصهٔ کاربر (سابقهٔ دانش فنی وی) ممکن است برای یک کاربر خاص، مربوط و یا نامربوط تلقّی شود. در تمام سیستمهای بازیابی اطّلاعات، چنانچه مدرک بازیابی شده در قضاوت کاربر، مورد توجه وی واقع گردد، آن مدرک به عنوان مدرک مربوط و در غیر این‌صورت، مدرک نامربوط شناخته می شود. عوامل بسیاری در قضاوت درباره عنصر «ربط» مؤثرند. از آنجا که عوامل بسیاری، قضاوت درباره ربط را با استفاده از روشهای پیچیده تعیین می کنند، یک سیستم بازیابی اطّلاعات نمی‌تواند به طور دقیق تمام مدارک مربوط را انتخاب نماید. بنابراین، سیستم باید روشهایی را بپذیرد که رتبه‌بندی مدارک را به ترتیب احتمال استفادهٔ کاربر از آنها آسان کند.

یکی از روشهای مناسب، محاسبهٔ همبستگی اصطلاحات، بر اساس فراوانی اصطلاحات هم آیند است. در صورت فرض تعامد بردارهای مدارک و اصطلاح، استفاده از ماتریس هم آیند می‌تواند یک عامل تنظیم‌کننده باشد. پژوهشگران این حوزه، در فرایند بازیابی، روشهای متفاوتی برای تشخیص همبستگی اصطلاحات ارائه کرده‌اند، از جمله می‌توان به تحلیل آماری جستجوها در مدارک مربوط و نامربوط به ترتیب همبستگیهای مثبت و منفی اصطلاحات، اشاره کرد [۵]. در پژوهش دیگری که از ماتریس اصطلاح هم آیند استفاده شد، مجموعهٔ اصلی بردارهای اصطلاح، از روشهای تحلیل عامل یا مقیاس چند بعدی بدست آمد [۷ و۶].

«کال» در پژوهش خود، طرحی ارائه کرد که با استفاده از آن می‌توان همبستگیهای میان اصطلاحات را بدون ماتریس اصطلاح هم آیند، ادغام نمود [۸].

در یک سیستم بازیابی اطّلاعات، معمول است که یک مدرک به وسیلهٔ کلید واژه‌ها یا واژه‌های موضوعی نمایانده شود. کلید واژه‌ها معمولاً در فرایند نمایه‌سازی، از متن یا چکیدهٔ مدرک استخراج می شوند.

علاوه بر گزینش اصطلاحات برای بازنمون مدارک، معمولاً به هر اصطلاح وزنی می‌دهند تا اهمیّت آن اصطلاح خاص را در مدرک نشان دهد. بنابراین، در طراحی استراتژیهای جستجو می‌توان ماتریس مدرک ـ اصطلاح را به وجود آورد، به نحوی که عنصر (i , r) این ماتریس، متناظر با وزن اصطلاح i در مدرک r باشد [۹]. در این ماتریس، عنصر d_{ir} را به عنوان مؤلفهٔ i بردار متناظر با مدرک r در نظر می‌گیرند. هنگام جستجو، سیستم، بردار جستجو را به دست می‌آورد و با بردارهای مدارک بر اساس روش بیان تشابه میان بردارها منطبق می سازد [۴]. با در اختیار داشتن این ماتریس و با توجه به هدف رتبه‌بندی مدارک، روشهای گوناگونی برای مدلسازی بازیابی به کار می‌رود. یک روش که در سالهای اخیر به طور گسترده استفاده شده، مدلسازی مدارک و جستجوها بر اساس بردار است [۱۰ و ۹] و به آن مدل «فضای برداری»[1] گفته می شود. هر جستجو به صورت بردار نمایانده می شود و تشابه آن را با بردارهای مدارک در نظر می‌گیرند. هر چه بردار جستجو به بردار مدرک نزدیکتر باشد، به موضوع جستجو مربوط تر خواهد بود. مدل فضای برداری را «سالتون» پیشنهاد کرده است [۱۱ و ۱۰ و ۹].

مدل فضای برداری

مدل فضای برداری یکی از مدلهای بازیابی اطّلاعات است که در سطح وسیعی به کار می‌رود [۱۳ و ۱۲]. در این مدل، هر مقولهٔ اطّلاعاتی ـ شامل متون ذخیره شده و هر تقاضای اطّلاعاتی زبان طبیعی ـ به صورت مجموعهٔ بردارهایی از اصطلاحات نگهداری

1. Vector space model.

می‌شوند. به طور نظری، این اصطلاحات می‌توانند از واژگان کنترل شده انتخاب شوند. به خاطر وجود مشکلاتی در تهیهٔ این واژگان، اصطلاحات از متون استخراج می‌شوند. معمولاً برای کاهش اندازهٔ واژگان از ریشهٔ واژه‌ها استفاده می شود. همچنین معمولاً از واژه‌های بازدارنده نظیر the, of, an ,....صرف نظر می گردد. از تمام واژه‌های موجود در مدارک، یک مجموعه واژگان به وجود می آید. هر مدرک به صورت برداری از تمام واژگان نمایانده می شود. بعید است واژه‌هایی که فاقد بار معنایی هستند و به طور معمول در مدارک یافت می‌شوند، اطّلاعات مهّمی ارائه دهند، بنابراین می‌توان این واژه‌ها را برای سرعت دادن به پردازش، حذف کرد. واژه‌های تکراری که می‌توان از آنها چشم پوشید فهرست واژه‌های غیرمجاز را می‌سازند. در حذف واژه‌های غیر مجاز، باید دقّت زیاد به کار برده شود. برای مثال:

چنانچه واژه‌های غیر مجاز در جملهٔ: «to be or not to be» حذف شوند، این جمله غیر قابل بازیابی خواهد بود.

مدل فضای برداری، شیوه ای است برای نمایش مدارک از طریق واژه‌های موجود در آنها. این مدل، یک تکنیک استاندارد در بازیابی اطّلاعات است. بر اساس مدل فضای برداری، می‌توان تصمیم گرفت که کدام مدارک شبیه به یکدیگر و یا به کلیدواژه‌های جستجو شبیه هستند [۱۲].

بردار مربوط به هر مدرک (یا هر جستجو) دارای n مؤلّفه است. n برابر با تعداد اصطلاحات موجود در مجموعه مدارک است.

به هر یک از اصطلاحات هر مدرک، به طور خودکار وزنی اختصاص می‌یابد که بر فراوانی رخداد اصطلاح در کلّ مجموعهٔ مدرک و تعداد دفعات حضور یک اصطلاح در مدرک خاص مبتنی است. با افزایش فراوانی اصطلاح در یک مدرک، وزن آن اصطلاح درمدرک افزایش می یابد. برعکس، وقتی فراوانی اصطلاح در مجموعهٔ مدارک بیشتر باشد، این وزن کاهش می یابد.

به طور کلی، می‌توان مزیّتهای اصلی مدل فضایی برداری را چنین بیان نمود [۱۳]:

۱. طرح وزن دهی به اصطلاح در این مدل، عملکرد بازیابی را بهبود می بخشد.

۲. استراتژی تطبیق جزئی این مدل، بازیابی مدارکی را مجاز می شمارد که به شرایط جستجو نزدیک هستند.

۳. فرمول رتبه بندی کسینوسی آن، مدارک را بر طبق درجهٔ تشابهی که به موضوع جستجو دارند، مرتّب می کند.

۲-۱. وزن دهی به اصطلاح

چون اصطلاحات متفاوت، دارای اهمیت مختلفی در متن هستند، از یک نشانگر مهم، یعنی «وزن اصطلاح» استفاده می‌شود که همراه هر اصطلاح است [۱۵و ۱۴]. به اصطلاحات مهمتر، وزن بیشتر اختصاص می یابد. برای وزن دادن به یک اصطلاح، از تعداد رویداد یک اصطلاح (فراوانی آن یا tf استفاده می‌شود). اهمیت اصطلاح، مستقل از بستر جستجو نیست. برای مثال، اصطلاح "دریاچه" در مجموعهٔ مقالات دریاچهٔ خزر اهمیت زیادی ندارد، اما در زمان جستجو در مقالات پیرامون کویر و صحرا، این اصطلاح می‌تواند بسیار مهم باشد. این امر نشان دهندهٔ این موضوع است که مدارکی که یک اصطلاح در آن مکرراً وجود دارد، ممکن است اهمیت کمتری داشته باشد. از این رو از «فراوانی مدرک معکوس» یا «عامل idf» همراه با وزن اصطلاح استفاده می‌شود. فراوانی اصطلاح معکوسِ برای محاسبهٔ اهمیت واژه‌های نادر نسبت به واژه‌های معمولی به وجود آمده است. فراوانی اصطلاح معکوس واژهٔ i به وسیلهٔ فرمول زیر محاسبه می شود.

$$idf_i = \log \frac{N}{n_i}$$

در اینجا N تعداد کلّ مدارک و n_i تعداد مدارک حاوی واژهٔ i است.

معمولاً در یک مدرک طولانی به طور مکرر از اصطلاح خاصی استفاده می شود. عامل فراوانی اصطلاح ممکن است برای مدرک طولانی، بزرگ باشد. همچنین مدرک طولانی، شامل بسیاری از اصطلاحات گوناگون و متمایز است. این امر باعث افزایش تعداد تطبیق واژه‌های یک جستجو و مدرک طولانی می شود و به طور ناعادلانه‌ای احتمال

بازیابی این مدرک را نسبت به مدارک کوتاه تر افزایش می‌دهد. برای جبران این اثر، معمولاً وزنهای اصطلاح را نرمال‌سازی می‌کنند. نرمال‌سازی فراوانی اصطلاح، یکی از اصلی ترین مباحث در بازیابی اطّلاعات طی سالهای اخیر است.

فراوانی اصطلاح به طول مدرک بستگی دارد. بنابراین، در یک سیستم بازیابی اطّلاعات، نیازمند یکنواخت کردن طول مدرک با تکنیک نرمال‌سازی فراوانی اصطلاح هستیم. بیشتر روشهای نرمال‌سازی فراوانی اصطلاح، در بازیابی اطّلاعات از پارامترها استفاده می‌کنند. تنظیم این پارامترها، می‌تواند به تغییرات مهمی در میزان دقت و بازیافت منجر گردد. یکی از پارامترهای مهم، وزن است.

نرمال‌سازی کسینوسی یکی از روشهای موثر نرمال‌سازی است. هر بردار مدرک به طول اقلیدسی آن تقسیم می شود، $\sqrt{w_1^2 + w_2^2 + ... + w_n^2}$ در اینجا، w_i وزن $tf \times idf$ اصطلاح i در مدرک است. وزن نهایی برای یک اصطلاح به صورت زیر محاسبه می گردد.

$$\frac{\text{وزن} \quad tf \times idf}{\text{طول اقلیدسی بردار مدرک}}$$

وزن اصطلاحی که در یک مدرک موجود نباشد را صفر در نظر می‌گیرند [۱].

باتوجه به نکات فوق می‌توان چنین گفت که تمام واژه‌های موجود در مدرک اهمیت یکسانی ندارند. یک واژه اگر دارای شرایط زیر باشد، به احتمال زیاد به مدرک d_1 بسیار مربوط خواهد بود:

الف) تکرار آن در سایر مدارک کم باشد.

ب) تکرار آن در مدرک d_1 بالا باشد.

۲-۲. رتبه‌بندی مدارک

تابع حاصل‌ضرب داخلی برداری می‌تواند برای یافتن همپوشانی واژگان میان هر دو بردار متن استفاده گردد. جستجوی زبان طبیعی که کاربر انجام می‌دهد، به بردار وزن داری

تبدیل می‌شود و با استفاده از تابع حاصل‌ضرب داخلی، تشابه عددی میان بردار جستجو و بردار هر مدرک در مجموعه محاسبه می گردد. با در نظر گرفتن بردار جستجوی Q و نمایش برداری مدرکi به صورت D_i، تشابه میان جستجو و مدرک به صورت زیر محاسبه می شود [۱]:

$$Sim(Q, D_i) = \sum_{t_j \in \text{اصطلاحات مشترک}} q_j \times d_{ij}$$

در اینجا t_j اصطلاحی است که در جستجو و مدرک، ظاهر شده و q_j وزن اصطلاح t_j در جستجو و t_{ij} وزن آن در مدرک i است. تمام اصطلاحات t_j که هم در جستجو و هم در مدرک وجود دارند با هم جمع می‌شوند. تشابه حاصل‌ضرب داخلی فهرستی از مدارک رتبه‌بندی شده با توجه به میزان استفاده آنها ارائه می دهد.

به طور معمول، کاربر جستجویی را در پایگاههای اطّلاعاتی وارد می کند. جستجو با تمام مدارک با اندازه‌گیری تشابه مقایسه می شود. مدارک به ترتیب نزولیِ تشابهی که با اصطلاح جستجو دارند، به کاربر ارائه می شود.

۳-۲. محاسبهٔ تشابه

روشهای مختلفی برای اندازه گیری تشابه میان دو مدرک، یا تشابه یک مدرک با یک جستجو وجود دارد. اندازه‌گیری کسینوسی، یک روش بسیار معمول اندازه‌گیری تشابه است، که در این روش کسینوس زاویهٔ بین مدرک و جستجو اندازه گیری می شود. با اندازه‌گیری تشابه، مجموعه ای از مدارک را می‌توان با جستجو مقایسه کرد و آنگاه مربوط ترین مدرک را بازیابی نمود [۱۲].

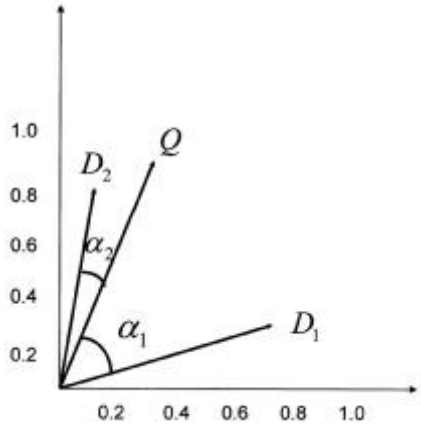
اندازه گیری کسینوسی:

برای دو بردار d و q، تشابه کسینوسی بین d و q به صورت زیر محاسبه می شود:

$$Sim(d,q)=\frac{d\times q}{|d||q|} \qquad (۱)$$

در اینجا، $d\times q$ حاصل‌ضرب برداری d و q است که با ضرب کردن فرایندهای متناظر در هم، محاسبه می شود.

اندازه گیری کسینوسی، زاویهٔ بین بردارها را در فضای چند بعدی محاسبه می کند. برای نمونه، شکل یک را در نظر بگیرید.



شکل ۱. نمایش برداری دو مدرک و یک جستجو

با توجه به این شکل، تشابه میان مدرک D_1 و جستجوی Q، برابر با کسینوس زاویهٔ بین دو بردار، یعنی a_1 است

$$Sim(D_1,Q)=Cos(a_1)$$

و همین‌طور داریم:

$$Sim(D_2,Q)=Cos(a_2)$$

۲-۴. پیاده سازی مدل فضای برداری

برای مدل‌برداری، وزن $w_{i,q}$ مربوط به زوج $(k_i و d_j)$ مثبت و غیر دودویی است. علاوه بر آن، واژه‌های موضوعی در جستجو نیز وزن دار می باشند. فرض کنید $w_{i,q}$ وزن

مربوط به زوج $[k_i,q]$ است که در آن $w_{i,q} \geq 0$ است. سپس، بردار جستجوی $\vec{q}$ به صورت $\vec{q} = (w_{1,q}, w_{2,q}, ..., w_{t,q})$ تعریف می شود که در آن t تعداد کلّ واژه‌های موضوعی در سیستم است. همانطور که قبلاً گفته شد، برای مدرک d_j ، بردار آن به صورت $\vec{d}_j = (w_{1,j}, w_{2,j}, ..., w_{t,j})$ نمایش داده می شود.

بنابراین، مدرک d_j و جستجوی کاربر، مثل q ، به صورت بردارهای t بعدی نمایش داده می‌شود. در مدل برداری پیشنهاد می‌شود برای ارزیابی (سنجش) درجهٔ تشابه مدرک d_j با ملاحظهٔ جستجوی q ، از همبستگی میان بردارهای $\vec{d}_j$ و $\vec{q}$ استفاده شود. این همبستگی را می‌توان تعیین کمیّت کرد. برای مثال، می‌توان از کسینوس زاویهٔ میان این دو بردار به صورت زیر استفاده کرد [۲].

$$\frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w^2{}_{i,j}} \times \sqrt{\sum_{i=1}^{t} w^2{}_{i,q}}} = \frac{\vec{d}_j \bullet \vec{q}}{|\vec{d}_j| \times |\vec{q}|} \quad \mathrm{Sim}(d_j,q)$$

در این جا $|\vec{d}_j|$ و $|\vec{q}|$ نرم بردارهای مدرک و جستجو هستند. عامل (فاکتور) $|\vec{q}|$ تأثیری بر رتبه بندی (یعنی ترتیب مدارک) ندارد، زیرا این عامل برای تمام مدارک، یکسان است. عامل $|\vec{d}_j|$ امکان نرمال‌سازی مدارک را فراهم می‌کند. در اینجا وزن از فرمول زیر به دست آمده است:

$$w_{i,j} = f_{i,j} \times \log \frac{N}{n_i}$$

از آن جا که $w_{i,j} \geq 0$ و $w_{i,q} \geq 0$ است، مقدار $\mathrm{Sim}(q,d_j)$ از ۰ تا ۱ متغیّر است. بنابراین، به جای پیش‌بینی مربوط یا نامربوط بودن یک مدرک، مدل برداری، مدارک را بر اساس درجهٔ تشابه آنها نسبت به جستجو رتبه‌بندی می‌نماید. ممکن است یک مدرک، حتّی در حالت تطبیق جزئی با جستجو بازیابی گردد. برای مثال، می‌توان آستانه ای را برای $\mathrm{Sim}(d_j,q)$ تعیین کرد و مدارکی را که درجهٔ تشابه آنها بیش از آستانه است، بازیابی نمود [۱۳].

نحوهٔ محاسبهٔ رتبه بندی مدارک

برای تبیین مسئله، از مثال ساده‌ای برای نشان دادن نحوهٔ ساخت بردار استفاده می‌کنیم:

مثال: مدارک و جستجوی زیر را در نظر بگیرید:

d_1: speech recognition and image processing and signal processing

d_2: speech models and image processing

q: image models

برای جستجو و هر یک از مدارک، جدولهایی شامل اصطلاحات موجود در هر رکورد ساخته می شود.

برای هر یک از اصطلاحات مستقل، یک مؤلفهٔ در جدول در نظر گرفته شده و فراوانی اصطلاح نیز مشخص گردیده است.

d_1: speech recognition and image processing and signal processing

speech	recognition	and	Image	processing	Signal
1	1	2	1	2	1

d_2 : speech models and image processing

speech	Models	and	image	Processing
1	1	1	1	1

q: image models

image	Models
1	1

در این مثال، تمام واژه‌ها برای تهیهٔ مجموعه واژگان به کار رفته‌اند.

واژگان ایجاد شده شامل تمام واژه‌هایی است که در مدارک به کار رفته‌اند:

speech, recognition, and, image, processing, signal, models.

درمدل فضای برداری، واژه‌های موجود در واژگان، مرتب سازی می شود:

and, image, models, processing, recognition, signal, speech.

بنابراین در این مثال بردارهای مدارک و جستجو دارای ۷ مؤلفه (به تعداد واژگان) بوده و بردار حاصل، هفت بُعدی خواهد بود.

بردارمدرک d_1 باتوجه به فراوانی اصطلاحات موجود در آن به صورت زیر نشان داده می شود:

d_1: speech recognition and image processing and signal processing

and	image	Models	processing	recognition	signal	Speech
2	1	0	2	1	1	1

$$Vec(d_1) = \vec{d}_1 = (2,1,0,2,1,1,1)$$

و بردار مدرک d_2 به صورت زیر به دست می آید:

d_2 : speech models and image processing

and	image	Models	processing	recognition	signal	speech
1	1	1	1	0	0	1

$$Vec(d_2) = \vec{d}_2 = (1,1,1,1,0,0,1)$$

بردار جستجو را می‌توان مانند بردارهای مدارک به وجود آورد:

q: image models

and	image	Models	processing	recognition	signal	speech
0	1	1	0	0	0	0

$$Vec \quad (q) = \vec{q} = (0,1,1,0,0,0,0)$$

رتبه بندی مدارک موجود نسبت به جستجوی q به صورت زیر محاسبه می شود:

ابتدا ضریب تشابه هر یک از بردارهای مدارک d_1 و d_2 با بردار جستجوی q به شیوه زیر محاسبه می‌شود:

$$d_1 = (2,1,0,2,1,1,1) \quad and \quad q = (0,1,1,0,0,0,0)$$

$$d_1 \times q = (2\times 0 + 1\times 1 + 0\times 1 + 2\times 0 + 1\times 0 + 1\times 0 + 1\times 0 = 1$$

$$|d_1| = \sqrt{(2^2 + 1^2 + 0^2 + 2^2 + 1^2 + 1^2 + 1^2)} = \sqrt{12}$$

$$|q| = \sqrt{(0^2 + 1^2 + 1^2 + 0^2 + 0^2 + 0^2 + 0^2)} = \sqrt{2}$$

$$Sim(d_1,q)=\frac{d_1\times q}{|d_1||q|}=\frac{1}{\sqrt{12}\sqrt{2}}=0.204$$

ضریب تشابه مدرک d_2 با جستجوی q به ترتیب زیر به دست می آید.

$$Sim(d_2,q)=\frac{d_2\times q}{|d_2||q|}=\frac{2}{\sqrt{5}\sqrt{2}}=\frac{1}{\sqrt{10}}=0.632$$

با توجه به مقادیر به دست آمده برای ضریب تشابه، از آنجا که ضریب تشابه مدرک d_2 با جستجوی q 0.632 و ضریب تشابه مدرک d_1 با جستجوی q برابر با 0.204 است، نتیجه می‌شود که مدرک d_2، مدرک مربوط تری نسبت به مدرک d_1 برای جستجوی q است. بنابراین، رتبه‌بندی به صورت d_2 و d_1 است.

نتیجهٔ این محاسبات نیز با نگاه اجمالی به مدارک و جستجو قابل تأیید است.

نتیجه گیری

در این مقاله، از میان استراتژیهای مختلف بازیابی اطّلاعات، مدل فضای برداری به عنوان یکی از معتبرترین تکنیکهای بازیابی به طور ساده بیان و نشان داده شد که مدل‌برداری یک استراتژی رتبه‌بندی است که با مجموعه‌های عمومی بهبودپذیر است. این استراتژی مجموعه جوابهای رتبه‌بندی شده‌ای تولید می‌کند که بهبود آنها بدون بسط جستجو یا بازخورد میزان ربط در چارچوب مدل برداری، مشکل است. در رتبه‌بندی، روشهای بسیار مختلفی با مدل برداری مقایسه شده است، امّا به طور کلّی به نظر می‌رسد مدل برداری، یا برتر بوده و یا تقریباً به خوبی سایر روشهای موجود عمل می نماید. به علاوه، مدل برداری، آسان و سریع است. با توجّه به این دلایل، مدل برداری یک مدل بازیابی معتبر است.

منابع

[1] Salton, G. (1989) Automatic Text Processing – The Transformation, Analysis and Retrieval of Information by Computer, Addison – Wesley Publishing Co., Reading, MA, 1989.

[2] Salton, G. (1991) Developments in Automatic Text Retrieval, Science, 253, 974-980, August.

[3] Tai, X., Ren, F. Kita, K. (2001) An Information Retrieval Model based on Vector Space Method by Supervised Learning, Information Processing & Management.

[4] Raghavan, V.V., Wony, S.K.M. (1986) Critical Analysis of Vector Space Model for Information Retrieval; Journal of the American Society for information Science.

[5] Raghavan, V.V., Yu, C.T. (1979) Experiments on the Determination of the Relationships Between Terms. ACM Transactions on Database Systems no. 4. pp.240 – 260.

[6] Katter, R.v. (1967) A Study of Document Representations: Multidimension Scaling of Index Terms. SDC – Final Report.

[7] Switzer, P. (1964) Vector Images in Information Retrieval. Proceedings of the Symposium on Statistical Association Methods for Mechanical Documentation, Wash. D.C. (NBS Misc. Publ. 269, 1965) Stevens, M.E., Heilprin, L., Guiliano, V.E (eds.). pp. 163 – 171.

[8] Koll, M. (1979) Weird, An Approach to Concept – based Information Retrieval. ACM – SIGIR Forum, vol XIII, no. 4, (spring 1979), pp. 32- 50.

[9] Salton, G., McGill, M.J. (1983) Introduction to Modern Information Retrieval. McGraw hill, New York.

[10] Salton, G. (1971) The SMART Retrieval System – Experiments in Automatic Document Processing. Prentice – Hall, Englewood Cliffs, New Jersey.

[11] Salton, G. (1983) Dynamic Information and Library Processing. Prentice – Hall, Englewood Cliffs, New Jersey.

[۱۲]. گراسمن، دیوید و افیر فریدر. «بازیابی اطّلاعات، الگوریتمها و روشهای اکتشافی» ترجمه جعفرمهراد و سارا کلینی، انتشارات کتابخانه رایانه‌ای، کتابخانه منطقه ای علوم و تکنولوژی، ۱۳۸۴.

[13] Baeza - Yates, R. Ribeiro- Neto, B, Modern information Retrieval, Addison Wesley, 1999.

[14] Salton, G., Yang, C.G., Yu, C.T. (1975) A Theory of Term Importance in Automatic Text Analysis, Journal of the ASIS, 26:1, 33-44.

[15] Salton, G. (1988) Buckley, C., Term weighting Approaches in Automatic Text Retrieval, Information Processing and Management, 24:5, 513-523.

[۱۶]. داورپناه، محمدرضا (۱۳۸۴). «ضرورتهای نوین بازنگری در ذخیره و بازیابی اطّلاعات». **کتابداری و اطّلاع رسانی**، جلد ۸ ، شمارهٔ ۳ ، پاییز ۱۳۸۴، ص ۶۷-۸۸.

[۱۷]. چاودری، جی جی، (۱۳۷۹). «پژوهش دربارهٔ اینترنت و بازیابی اطّلاعات». ترجمهٔ مهدی خادمیان، **کتابداری و اطّلاع رسانی**، جلد ۳ ، شماره ۳ ، پاییز ۱۳۷۹، ص ۱۳۳-۱۶۲.