نشریه علمی- ترویجی محاسبات نرم

شمارهٔ ششم/ پاییز و زمستان ۹۳/ صفحهٔ ۳۱-۴۱

ارسال مقاله: ۹۳/۸/٦

پذیرش مقاله: ۹٤/۲/۷

بهبودی در سیستم‌های پیشنهادگر خبره با استفاده از بسط پرسش و مدل فضای برداری

احسان پرنور[۱]، جلال رضایی‌نور[۲]

[۱] دانشجوی کارشناسی ارشد فناوری اطلاعات، دانشکده فنی و مهندسی ـ دانشگاه قم ـ قم- ایران

e.pournoor@hotmail.com

[۲] استادیار، گروه مهندسی صنایع، دانشکده فنی و مهندسی ـ دانشگاه قم- قم ـ ایران

J.rezaee@qom.ac.ir

چکیده: با توجه به افزایش میزان حجم داده‌های موجود روی وب، یافتن دانش مورد نیاز از میان حجم انبوه داده‌ها امری بسیار مشکل می‌باشد. سیستم‌های پیشنهادگر دانش، فروم‌های آنلاین و سیستم‌های پاسخ به پرسش جهت آسان کردن راه دسترسی به دانش مورد نیاز و پاسخگویی به نیاز اطلاعاتی کاربران بوجود آمده‌اند. سیستم‌های پاسخ به سوال با ایده‌ی پاسخ‌دهی کوتاه و مفید با استفاده از مخازن دانشی ثبت شده، دسترسی به دانش مورد نیاز را بهبود داده‌اند. اما با توجه به ماهیت دانش، انتقال آن مشکل‌تر از انتقال اطلاعات است. سیستم‌های پیشنهادگر خبره با پیشنهاد افراد متخصص علاوه بر انتقال اطلاعات، باعث انتقال تجارب و درک در مورد موضوع می‌شوند. این سیستم‌ها از تحلیل محتوایی پروفایل متنی و سوابق متخصصین، تحلیل ارتباطات متخصصین و ترکیب این دو روش برای یافتن متخصصین استفاده می‌کنند. در این مقاله از داده‌های رزومه متخصصین (اساتید دانشگاه) استفاده شده است و با استفاده از مدل فضای برداری و استفاده از تکنیک بسط پرسش، مدلی جدید برای خبره‌یابی ارائه شده که با تحلیل محتوایی اسناد، خبره‌های دارای تخصص لازم را پیشنهاد می‌کند. نتایج شبیه‌سازی‌ها بیانگر این موضوع است که مدل ارائه شده دقت بالاتری نسبت به مدل فضای برداری دارد.

واژه‌های کلیدی: سیستم‌های پیشنهادگر، فروم‌های آنلاین، خبره‌یابی، مدل فضای برداری

۱. مقدمه

در شرایطی که هر روز بر حجم میزان اطلاعات موجود در وب افزوده می‌شود، وجود سیستمی که بتواند از میان حجم عظیمی از اطلاعات، اطلاعات مرتبط و مورد نیاز را بنا به نیاز و ویژگی‌های هر کاربر به وی پیشنهاد کند، به موضوعی بسیار مهم تبدیل شده است. موتورهای جستجو با بازیابی اسناد مرتبط به پرسش کاربر، یافتن مطلب مورد نیاز را فراهم کرده‌اند. با این وجود، هنوز میزان اسناد بازیابی شده زیاد است و پیدا کردن پاسخ از بین اسناد بازیابی شده دشوار است. سیستم‌های پیشنهادگر با پیشنهاد فیلم، موسیقی، کالا و ... راه پیدا کردن کالای مورد نیاز را آسان می‌کنند. هدف سیستم‌های پیشنهادگر در وهله اول شناخت ویژگی‌های کاربر و پیشنهاد کالای مورد نیاز به اوست. در وهله‌ی دوم، وظیفه‌ی سیستم‌های پیشنهادگر، پیشنهاد چیزهایی است که کاربر از وجود آنها بی‌اطلاع است. کاربران فضای مجازی، در مواجهه با نیاز اطلاعاتی، برای یافتن پاسخ به نیاز اطلاعاتی خود با حجم بزرگی از اطلاعات بدون ساختار مواجه هستند که امکان یافتن پاسخ صحیح در زمان اندک بسیار مشکل می‌باشد. لذا نیاز به سیستم‌هایی می‌باشد که دسترسی به این اطلاعات را تسهیل کنند. سیستم‌های پیشنهادگر دانش، نوعی از سیستم‌های پیشنهادگر هستند که با درک نیاز کاربر، دسترسی به دانش مورد نیاز را در زمان اندک فراهم می‌کنند.

سیستم‌های خبره‌یاب نوع دیگری از سیستم‌های پیشنهادگر هستند که با پیشنهاد افراد خبره سعی بر ارائه مفید و مؤثر مطالب با استفاده از انتقال فهم و درک در یک موضوع را دارند.

این سیستم‌ها با پیشنهاد متخصصین به عنوان منابع دانشی پدیدآورنده‌ی دانش، امکان انتقال سریع و دقیق دانش را فراهم کرده‌اند.

در این مقاله، مدلی جدید برای سیستم‌های پیشنهادگر خبره ارائه شده است که با استفاده از مدل فضای برداری به عنوان مدلی مرجع برای سیستم‌های بازیابی اطلاعات و استفاده از تکنیک بسط پرسش، متخصصین را بر اساس میزان مشابهت به پرس‌وجوی کاربر پیشنهاد می‌کند.

۲. مروری بر کارهای گذشته

سیستم‌های خبره‌یاب سیستم‌هایی هستند که وظیفه‌ی آنها پیشنهاد متخصصین در حوزه‌ی یک موضوع می‌باشد. خبره‌یابی سنتی بیشتر در داخل سازمانها وجود داشت که در آنها اسناد و مخازن دانشی وجود داشت ودانش به صورت اسناد ساختاریافته ثبت شده بود و کیفیت اطلاعات بالا بود. در مقایسه با خبره‌یاب‌های داخل سازمان، در فروم‌های دانشی آنلاین ساختار دانش جهانی وجود ندارد. افراد به اندازه‌ای که در این فروم‌ها شرکت دارند و به اندازه‌ای که نیاز دانشی خود را حل کنند به پرسش و پاسخ پرداخته‌اند. دانشی که در فروم‌های آنلاین وجود دارد عمدتا کیفیت بالایی ندارد و این به نوبه‌ی خود عملکرد سیستم‌های خبره‌یاب را دچار مشکل می‌کند. کیفیت اطلاعات به میزان بالایی به خوب بودن و مفید بودن سیستم‌های فناوری اطلاعات بستگی دارد [۱]. سیستم‌های خبره‌یاب از روش‌های زیر استفاده می‌کنند: روش محتوایی، روش تحلیل شبکه و روش‌های ترکیبی.

- **خبره‌یابی بر اساس تحلیل محتوایی**

در این تکنیک برای پیدا کردن متخصص، از روش‌های بازیابی اطلاعات برای یافتن مشابهت بین اسناد و پرس‌وجو استفاده می‌شود. خبره‌ها بر اساس مشابهتی که بین نوشتجات آنها وجود دارند شناخته می‌شوند. برخی از محققین از روش پروفایلینگ برای یافتن متخصصین خود استفاده کرده‌اند. سیستم‌های yellowpages.com، guru.com، 88owls.com از جمله سیستم‌هایی هستند که برای خبره‌یابی از پروفایلینگ استفاده می‌کنند[۲]. به این ترتیب که برای هر متخصص پروفایلی شامل اطلاعات شخصی آنها، مطالب مورد علاقه‌ی آنها که با کنترل صفحات مشاهده‌شده‌ی آنها جمع‌آوری می‌شود، ایجاد می‌شود و پس از آن از این اطلاعات به عنوان حوزه تخصصی فرد و برای محاسبه مشابهت حوزه دانشی آنها استفاده می‌شود. این پروفایل‌ها بر اساس کلماتی که در آنها استفاده شده است توصیف می‌شوند. برخی دیگر از نوشتجات متخصصین در فروم‌ها و سوالات و پاسخ‌های آنها استفاده می‌کنند و با استفاده از کلمات مورد استفاده در این نوشته‌ها

حوزه تخصصی فرد را تشخیص می‌دهند. سیستم‌های خبره-یاب زیادی وجود دارند که با استفاده از روشهای متن کاوی این مطالب را کاوش کرده و به صورت اتوماتیک تخصص افراد را مشخص می‌کنند[۳,٤,٥]. این مطالعات بر این باورند که درجه تخصص یک فرد به اسناد منتشر شده از او بستگی دارد [٦,٧,٨]. همچنین پاسخ‌های افراد را که به عنوان پاسخ-های منتخب و درست انتخاب شده‌اند جمع‌آوری کرده و از تاریخچه پاسخ‌های منتخب برای محاسبه میزان مشابهت استفاده می‌کنند. در [٧,٨] بالوگ و همکاران مدلی برای تعیین خبره‌ها با استفاده از مشابهت پرس‌وجو و اسناد خبره ارائه کرده‌اند. همچنین در [۳] کرولیچ و همکارانش سیستم ContactFinder را که از تاریخچه پیغام‌های افراد برای شناسایی متخصصین استفاده می‌کرد را توسعه داده‌اند. عمده پژوهش‌هایی که از تحلیل محتوایی برای یافتن خبره‌ها استفاده کرده‌اند از مدل‌های بازیابی اطلاعات برای شناسایی افراد متخصص استفاده کرده‌اند. معروف‌ترین مدل بازیابی اطلاعات که در کل نتایج بهتری نسبت به سایر مدل‌ها دارد، مدل فضای برداری است[۹]. در مدل فضای برداری، برای سنجش میزان ربط اسناد و نیاز اطلاعاتی کاربر، سیستم اسناد موجود و پرسش کاربر را در فضای چند بعدی مدل‌سازی می‌کند. در نتیجه برای سنجش میزان شباهت میان بردار پرسش و بردار هر سند می‌توان از زاویه‌ای که این دو بردارها با هم می‌سازند، استفاده کرد. در این مدل تمام اسناد و همچنین پرس‌وجوهای کاربران به عنوان برداری در فضای کلمات در نظر گرفته می‌شوند. برای مثال اگر t کلمه‌ی مجزا در کل مجموعه‌ی ما وجود داشته باشد، یک فضای t بعدی از کلمات خواهیم داشت. تمامی اسناد و همچنین پرس‌وجوی کاربران بر اساس سهمی که از این کلمات دارند، برداری در این فضا خواهند بود. مدل فضای برداری میزان مشابهت سند و پرس‌وجو را ارزیابی می‌کند و برای سنجش میزان مشابهت از وزن کلمات در اسناد استفاده می‌کند. معادله (۱) روش محاسبه میزان شباهت سند d_j را به پرس‌وجوی q و در فضای t بعدی (t تعداد کل کلمات کلیدی موجود در مجموعه است) نشان می‌دهد. در این معادله $w_{i,j}$ وزن کلمه i را در سند d_j و $w_{i,q}$ وزن کلمه i را پرس‌وجوی q نشان می‌دهد.

$$sim(d_j,q)=\frac{\vec{d}_j \bullet \vec{q}}{\left|\vec{d}_j\right| \times \left|\vec{q}\right|}=\frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{j=1}^{t} w_{i,q}^2}} \qquad (۱)$$

وزن‌دهی کلمات در اسناد نیز با استفاده از رابطه‌ی (۲) انجام می‌شود:

$$w_{i,j}=f_{i,j} \times \log \frac{N}{n_i} \qquad (۲)$$

در این رابطه $f_{i,j}$ تعداد تکرار کلمه i را در سند d_j و n_i نشان‌دهنده‌ی تعداد اسنادی است که کلمه‌ی i در آنها حضور داشته است. N نیز تعداد کل اسناد موجود در مجموعه می‌باشد. همچنین وزن‌دهی کلمات موجود در پرس‌وجو در مدل فضای برداری نیز با استفاده از معادله (۳) انجام می‌شود:

$$w_{i,q}=\left(0.5+\frac{0.5 freq_{i,q}}{\max_l freq_{l,q}}\right) \times \log \frac{N}{n_i} \qquad (۳)$$

- **خبره‌یابی بر اساس روش‌های تحلیل شبکه**

اخیرا برای بهبود الگوریتم‌های خبره‌یابی از روش‌های تحلیل گراف استفاده شده است. این مطالعات نشان داده‌اند در محیط‌هایی مثل شبکه‌های اجتماعی که اشخاص با یکدیگر دارای ارتباطات اجتماعی می‌باشند، استفاده از الگوریتم‌های بر اساس گراف نتایج بهتری داشته‌اند. آنها ادعا می‌کنند که انتقال دانش فردی از ارتباطات شخصی و غیر رسمی ناشی می-شود[۱۰,۱۱,۱۲]. برخی از مطالعات از ایده‌ی زنجیره‌ی مارکو که ایده‌ی جدیدی در حوزه‌ی خبره‌یابی است استفاده کرده-اند[۱۳]. این مطالعات خبره‌یابی را در شبکه‌های اجتماعی بررسی کرده و از اطلاعات ارتباطی مثل ارتباطات دوستی، ارتباطات پیگیری مطالب، پرس‌وجو کننده- پاسخ دهنده، فرستنده – گیرنده ایمیل و در محیط‌های اجتماعی آکادمیک مثل مقاله‌ی مشترک داشتن، هم‌رشته بودن برای یافتن تخصص فرد از روی تخصص مرتبطین آنها استفاده کرده‌اند. مطالعات

نشان می‌دهند استفاده از اطلاعات پیرامونی افراد در فعالیت-های جستجوی دانش بسیار مهم است. در [۱٤] بیان شده است که ارتباط مثبتی بین تعداد پیام‌هایی که یک شخص و دیگران ارسال شده است و میزان شایستگی او وجود دارد. مطالعات سیستم‌های اطلاعاتی نشان می‌دهد اطلاعات پیرامونی افراد و نه اطلاعات محتوایی و اسنادی آنها در فعالیت‌های جستجوی دانش بسیار مؤثر است. در روش‌های تحلیل گراف عمدتا از الگوریتم‌های PageRank و HITS به عنوان روش‌های مرجع استفاده شده است. در [۱۵] شبکه ارتباطات فروم مباحثه‌ای براساس لینک‌های بین پرسش‌کننده و پاسخ‌دهنده مورد ارزیابی قرار گرفته شده است. آنها الگوریتم ExpertiseRank را بر اساس الگوریتم PageRank ارائه کرده و بر اساس اینکه چه کسی، چه کسی را کمک کرده است، خبره‌یابی را انجام داده‌اند. آنها بیان کرده‌اند که الگوریتم PageRank در مقایسه با سایر روش‌ها نتایج بهتری دارد. در [۱٦] با استفاده از ایجاد لینک ارتباطات ما بین فرستنده و گیرنده در شبکه‌ی ارسال ایمیل و ما بین پرسش کننده و پاسخ-دهنده در فروم Yahoo، شبکه اجتماعی تشکیل داده و با تعیین شناسه[1] برای هر فرد، میزان خبرگی هر فرد با استفاده ار الگوریتم HITS مشخص شده است.

- **خبره‌یابی بر اساس روش‌های ترکیبی**

در مطالعات اخیر از روش‌های ترکیبی برای خبره‌یابی استفاده شده است. در این مطالعات، هم از روش تحلیل محتوایی و هم از روشهای تحلیل شبکه استفاده شده است [۱۸,۱۷,۱۵]. از این روش، بیشتر در شبکه‌های اجتماعی که افراد هم دارای پروفایل و اطلاعات مکتوب هستند و هم دارای روابط با متخصصین دیگر می‌باشند، مورد استفاده قرار گرفته است. استفاده از روشهای ترکیبی نتایج بهتری را نسبت به روشهای پیشین داشته است. در [۱۹] با استفاده از تاریخچه نوشتجات کاربران و اطلاعات شخصی آنها و با ایجاد پروفایل دانش برای هر کاربر و تحلیل محتوایی آن امتیاز هر خبره را مشخص کرده و سپس با تحلیل لینک‌های پرسش و پاسخ و با استفاده از ارتباط پرس‌وجو و پرس‌وجوهای پیشین خبره‌یابی را انجام می‌دهند. آنها همچنین از پاسخ‌های منتخب افراد به عنوان بهترین پاسخ استفاده کرده و شهرت افراد را محاسبه کرده‌اند. در [۲] مدل ExpertRank را براساس الگوریتم PageRank ارائه کرده‌اند. آنها از ترکیب تحلیل محتوایی وتحلیل شبکه استفاده کرده‌اند و نتایج آنها از هر دو روش محتوایی خالص یا تحلیل شبکه خالص بهتر بوده است. روش آنها دارای ۳ مزیت بوده است: ۱- استفاده در فروم‌هایی که اطلاعات آتها از کیفیت بالایی برخوردار نیست ۲- رتبه دهی اتوماتیک خبره‌ها در محیط جستجو بر اساس پرس‌وجو ۳- استفاده از هر دو روش محتوایی و تحلیل شبکه. در [۲۰] سردیکو و همکارانش یک الگوریتم ارائه کردند که از هر دو روش بر پایه‌ی سند و ارتباطات غیرمستقیم سند – نویسنده استفاده می‌کرد. آنها همچنین از اطلاعات دیگری مثل لینک صفحات وب[2]، ساختار و بازخورد کاربران استفاده کرده‌اند. ارزیابی‌ها نشان می‌دهند که روش آنها از الگوریتم‌هایی که تنها اسناد را در نظر می‌گرفتند، بهتر بوده است. کمپل و همکاران در [۲۲,۲۱] دو الگوریتم بر پایه‌ی Hits را در حوزه‌ی ارتباطات ایمیل ارائه کردند. الگوریتم از هردوی محتوای آنها و الگوهای ارتباطی آنها استفاده کرده است تا متخصصین را رتبه‌بندی کند. ارزیابی‌ها نشان می‌دهد که نتایج آنها نسبت به روش بر پایه‌ی سند بهتر بوده است. فو و همکاران [۲۳] تلاش مشابهی برای برای ترکیب روش محتوایی و بر پایه‌ی شبکه اجتماعی برای خبره‌یابی انجام دادند. آنها یک شبکه اجتماعی از ارتباطات ایمیل و رویارویی افراد در صفحات وب ایجاد کردند. آنها بهبود عملکرد را در مقایسه با روش‌های پایه نشان دادند. در [۲٤] نیز از یک روش ترکیبی برای خبره‌یابی استفاده شده است. آنها از مقالات علمی برای تشخیص خبره‌ها استفاده کرده‌اند. آنها از روش‌های متن‌کاوی و تحلیل محتوایی برای رتبه‌بندی اولیه و پس از آن، با استفاده از الگوریتمی مشابه PageRank لینک ارتباطات را مورد تحلیل قرار داده‌اند.

در کل می‌توان فعالیت‌های انجام شده در زمینه خبره‌یابی در جدول (۱) خلاصه کرد.

---

[1]. Authority

[2]. Hyperlink

۳. روش پیشنهادی

- **مرحله اول**

در بازیابی اطلاعات به روش فازی از ماتریس همبستگی کلمات تزاروس[1] برای بهبود پرس‌وجوی کاربر استفاده می‌شود. ماتریس همبستگی کلمات، ماتریسی است مربعی که ابعاد آن را تمام کلمات مجزای موجود در مجموعه اسناد تشکیل می‌دهند. سپس بر اساس میزان حضور کلمات در اسناد مختلف میزان همبستگی بین آنها مشخص می‌گردد. فرض کنید مجموعه‌ی کلمات موجود در کل مجموعه اسناد شامل $\{t_1, t_2, t_3, \ldots, t_n\}$ باشند در اینصورت ماتریس همبستگی کلمات به صورت زیر خواهد بود:

$$\begin{array}{c} \\ t_1 \\ \vdots \\ t_n \end{array} \begin{array}{c} \begin{array}{ccc} t_1 & \cdots & t_n \end{array} \\ \begin{pmatrix} c_{1,1} & \cdots & c_{1,n} \\ \vdots & 1 & \vdots \\ c_{n,1} & \cdots & a_{n,n} \end{pmatrix} \end{array}$$

شکل (۱) ماتریس همبستگی کلمات

در این ماتریس مقادیر $C_{i,j}$ نشان دهنده میزان همبستگی بین دو کلمه t_i و t_j است و مقدار آن با استفاده از رابطه‌ی (٤) محاسبه می‌شود:

$$c_{i,l} = \frac{n_{i,l}}{n_i + n_l - n_{i,l}} \qquad (٤)$$

در رابطه‌ی بالا، n_i نشان دهنده‌ی تعداد اسنادی است که کلمه t_i در آن حضور دارد و به همین ترتیب n_l تعداد اسنادی را مشخص می‌کند که کلمه t_l در آن وجود داشته‌است. همچنین $n_{i,l}$ نشان دهنده تعداد اسنادی است که دو کلمه t_i و t_l با هم در آن حضور داشته‌اند. پس از محاسبه مقادیر همبستگی به ازای هر دو کلمه موجود در مجموعه ماتریس همبستگی نشان دهنده‌ی میزان همبستگی کلمات به هم خواهد بود. در تئوری فازی از ماتریس همبستگی کلمات برای بهبود پرس‌وجوی کاربر استفاده شده است. به این معنی که به ازای کلمات موجود در پرس‌وجوی کاربر، با استفاده از ماتریس همبستگی کلمات، کلماتی که همبستگی زیادی دارند به پرس‌وجو افزوده می‌شوند و این به نوبه‌ی خود باعث بازیابی اسنادی می‌شود که شامل کلمات همبسته می‌باشند و لذا باعث افزایش میزان دقت در بازیابی می‌شود. برای مثال در صورتی که عبارت پرس‌وجوی کاربر از کلمات x و y تشکیل شده باشد $(query = xy)$، و اگر کلمات $a=x$ و $b=y$ کلمات مترادف (همبسته‌ی قوی با این دو کلمه باشند) در اینصورت با استفاده از بسط پرس‌وجو عبارت پرسش به $query = axby$ تبدیل می‌شود، تا اسنادی که در آنها کلمات a و b وجود دارند نیز بازیابی شوند. نتایج پژوهش‌های پیشین بیانگر این موضوع هستند که رتبه‌بندی بازیابی با استفاده از بسط پرس‌وجو به مراتب بهتر است. در مدل ارائه شده، از این ایده بهره گرفته و ابتدا پرس‌وجوی کاربر بسط داده می‌شود. پس در مرحله‌ی اول پرسش کاربر را بسط داده و سپس وارد مرحله‌ی دوم می‌شویم.

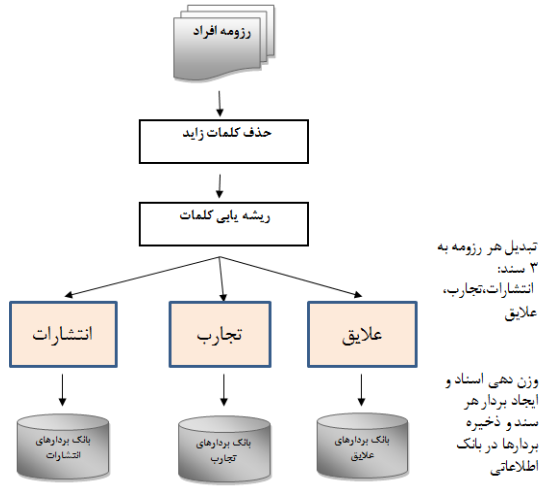
جدول (۱). مقایسه بین روشهای مختلف در سیستم‌های خبره‌یاب

روش	معایب	مزایا
تحلیل محتوایی (استفاده از مدل‌های بازیابی اطلاعات)	عدم استفاده از ارتباطات اجتماعی و ویژگی‌های رابطه‌ای	استفاده از داده‌های پروفایل و تخصصی افراد و نوشتجات آنها
تحلیل شبکه‌ای (استفاده از الگوریتم‌های تحلیل گراف)	استفاده از ارتباطات و شبکه‌های اجتماعی ما بین متخصین و تشخیص میزان خبرگی از طریق رفتارهای اجتماعی و اطلاعات پیرامونی افراد	عدم استفاده از داده‌های تخصصی افراد مثل نوشتجات و داده‌های پروفایل آنها
روش ترکیبی (استفاده مشترک از مدل‌های بازیابی اطلاعات و الگوریتم‌های تحلیل گراف)	------	استفاده از اطلاعات اجتماعی، بازخورد کاربران، داده‌های متنی و محتوایی افراد

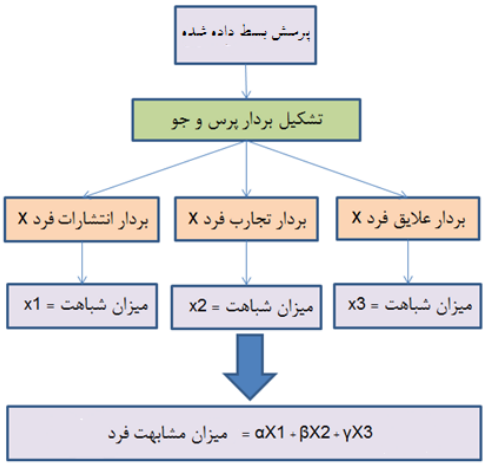
[1]. Thesaurus

- **مرحله‌ی دوم**

بدلیل اینکه اسناد استفاده شده در این پژوهش رزومه‌های متنی متخصصین هستند، در مدل ارائه شده، هدف رتبه‌بندی متخصصین بر اساس اولویت موارد ذکر شده در بخش‌های مختلف رزومه آنهاست. در این مدل بخش‌های مختلف رزومه در سه بخش انتشارات، علایق و تجربه‌ها دسته‌بندی شده و بر اساس اینکه چه کلماتی در این بخش‌ها حضور داشته است میزان شایستگی هر فرد را به جستجوی کاربر تعیین شده است. بخش انتشارات شامل مقالات علمی، کتاب‌ها و آنچه که متخصص منتشر کرده است، می‌باشد. بخش تجربه‌ها نیز شامل تجارب علمی، عملی، سوابق آموزشی و تدریس‌های متخصص می‌باشد و در بخش علایق اطلاعات علایق حوزه-های علمی و شخصی متخصص گنجانده می‌شود. ایده‌ی استفاده از این روش این است که برخی کلمات موجود در رزومه افراد دارای وزن اهمیت بیشتری نسبت به بقیه می‌باشند. برای مثال اهمیت کلماتی که در تجارب فرد وجود دارند از کلماتی که در بخش علایق هستند، بیشتر می‌باشد و آنچه که تخصص یک فرد را نشان می‌دهد بیشتر تجارب فرد است تا علایق او. لذا پس از بخش‌بندی رزومه‌ها به این سه بخش، هر بخش به صورت مجزا مورد پالایش قرار گرفته و در بانک اطلاعاتی مربوطه ذخیره شده است.



شکل(۲) نمایه سازی رزومه ها در مدل



شکل (۳) ترکیب وزنی بخش های رزومه در مدل ارائه شده

پس از ذخیره در بانک اطلاعاتی و تشکیل بردار هر بخش از سند رزومه در فضای کلمات مجموعه، با استفاده از مدل فضای برداری هر بخش به صورت مجزا مورد ارزیابی قرار گرفته و مشابهت هر بخش با پرس‌وجوی کاربر محاسبه شده و در نهایت میزان مشابهت (درجه اهمیت) بخش‌ها با هم ترکیب خطی شده و مشابهت نهایی تعیین می‌شود.

درصورتی که میزان مشابهت بردار علایق فرد x را با x_1 بردار تجارب او را با x_2 و بردار انتشارات فرد را با x_3 نشان دهیم در اینصورت آنچه که در ترکیب خطی این مشابهت‌ها بسیار مهم می‌باشد تعیین مقادیر α، β و γ می‌باشد. ترکیب وزنی بخش‌های مختلف رزومه در شکل (۳) نمایش داده شده است.

- **فرایند تحلیل سلسله مراتبی**

یکی از روش‌هایی که در تصمیم‌گیری برای وزن‌دهی معیارها و یا رتبه‌بندی و انتخاب گزینه‌ها می‌تواند مورد استفاده قرار بگیرد فرآیند تحلیل سلسله مراتبی[1] [۲۵] است. در این روش از مقایسات زوجی گزینه‌ها استفاده می‌شود و با ایجاد درخت سلسله مراتبی و بر مبنای آن مقایسه بین گزینه‌های کاندید انجام می‌شود. این مقایسات وزن هر کدام از فاکتورها را در راستای گزینه‌های رقیب مورد ارزیابی در تصمیم را نشان می-دهد. در این مقاله برای اولویت‌بندی گزینه‌های موجود (تجارب، انتشارات، علایق) از روش تحلیل سلسله مراتبی استفاده شده است.

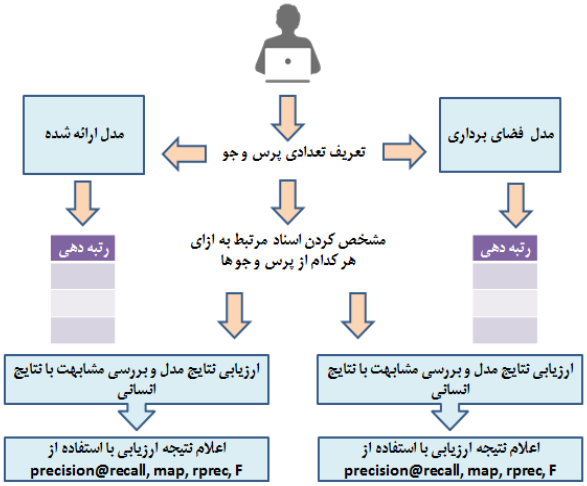
1. Analytical Hirarchy Process (AHP)

٤. شبیه‌سازی مدل

برای آزمایش نتایج پژوهش و مدل ارائه شده، برنامه کاربردی پریتی رنک (PrettyRank) شامل: نمایه‌ساز، ریشه‌یاب، جستجوگر و ارزیاب، را پیاده‌سازی کرده‌ایم. دقت برنامه کاربردی پریتی رنک با نرم افزار Lemur (برنامه مشابه و مرجع برای سیستم‌های بازیابی اطلاعات) مورد مقایسه و تایید قرار گرفته است. این برنامه داده‌ها (رزومه‌ها) را به صورت فایل-های متنی گرفته و مراحل نمایه‌سازی، ریشه‌یابی و ذخیره-سازی نمایه‌ها را در بانک‌های اطلاعاتی به صورت آفلاین انجام داده است. برنامه PrettyRank شامل بخش‌هایی برای دریافت پرس‌وجو و گزارش نتایج جستجو می‌باشد. همچنین دارای قابلیت اجرای پرس‌وجوی چندگانه بوده و فایل شامل پرس‌وجوها به همراه رزومه‌های مرتبط به این پرس‌وجوها را دریافت کرده و جستجو را انجام می‌دهد.

برای ارزیابی میزان دقت مدل ارائه شده، دقت مدل در مقایسه با یک مدل مرجع و با استفاده از ارزیابی انسانی سنجیده می-شود. همانطور که در شکل (٤) مشاهده می‌شود، ابتدا پرسش-هایی توسط عوامل انسانی طراحی شده و اسناد مرتبط با این پرسش‌ها نیز توسط متخصصین انسانی مشخص می‌شوند. مثلا برای پرسش "متخصص کامپیوتر آشنا به داده‌کاوی و بازیابی اطلاعات" متخصصین مربوطه توسط عوامل انسانی شناسایی می‌شوند. توجه به این نکته ضروری است که این نتایج دقیق-ترین نتایج می‌باشند و تنها توسط عوامل انسانی مشخص می-شوند. سپس بدون توجه به نتایج انسانی، مدل مرجع (مدل فضای برداری در این پژوهش) و مدل ارائه شده هر کدام به صورت مجزا نتایج خود را که متخصصین رتبه‌بندی شده است، برای هر پرسش ارائه می‌کنند. نتایج هر کدام از مدل مرجع و مدل ارائه شده به صورت مجزا با نتایج انسانی مقایسه می‌شوند. این مقایسه مشخص می‌کند که نتایج مدل مرجع و مدل پیشنهادی چقدر با تشخیص انسانی مشابهت دارد. در این مقایسه، معیارهای ارزیابی P@R، MAP، RPREC و Harmonic Mean اعدادی را به این دو مدل نسبت می‌دهند که میزان موفقیت هر کدام را در مقایسه با نتایج انسانی مشخص می‌کند. در نهایت میزان موفقیت دو مدل با هم مقایسه می‌شوند تا مشخص شود کدام مدل میزان موفقیت بالایی دارد. به این معنی که هر کدام از مقادیر P@R، MAP، RPREC و Harmonic Mean (F) برای مدل مرجع و مدل پیشنهادی با هم مقایسه می‌شوند و هر کدام که مقدار بالاتری داشت لذا نتایج دقیق‌تری دارد.

در سیستم‌های بازیابی مقدار دقت(Precision) را برای مقادیر بازخوردها (Recall) استاندارد (۵٪، ۱۰٪، ۲۰٪، ۳۰٪ و ٤۰٪ بازیابی و ...) می‌سنجند. معیارهای دیگر بازیابی مثل Harmonic Mean و بقیه معیارهای ترکیبی از دو مقدار Precision و Recall هستند.



شکل (۴) نحوه ارزیابی سیستم خبره‌یاب

برای آزمایش مدل ارائه شده، تعداد ۱۱۵ رزومه اساتید دانشگاه مورد استفاده قرار گرفته است. این رزومه‌ها نیز همگی انگلیسی بوده‌اند. رزومه‌های اساتید از دانشگاه‌های کشورهای انگلیس (دانشگاه منچستر، دانشگاه کمبریج، دانشگاه لستر[1])، کانادا (دانشگاه یورک)، استرالیا (دانشگاه بالارات)، ایران (دانشگاه بوعلی سینا، دانشگاه صنعتی شریف، دانشگاه قم، دانشگاه تهران، دانشگاه صنعتی همدان) بوده است. ابتدا بخش‌های مختلف این رزومه‌ها جدا شده و سه بخش اساسی آنها (انتشارات، علایق و تجارب) به صورت مجزا مورد پالایش قرار گرفته و در بانکهای اطلاعاتی ذخیره شده‌اند. در

[1]. University of Leicester

مرحله پالایش متنی رزومه‌ها، تمامی کلمات زاید حذف شده و کلمات توسط الگوریتم پورتر ریشه‌یابی شده‌اند.

برای این آزمایش، تعداد ۵۳ پرس‌وجو مشخص شده و رزومه‌های مرتبط به آنها توسط عوامل انسانی متخصص مشخص شده است. برنامه توسعه داده شده برای اجرای آزمایش، پرس‌وجوهای مشخص شده و لیست رزومه‌های مرتبط به آنها را گرفته، پرس‌وجوها را بسط داده و با استفاده از مدل فضای برداری هر بخش را به صورت مجزا رتبه‌دهی کرده و سه رتبه مربوط به هر رزومه را با هم ترکیب خطی کرده و نتیجه‌ی نهایی مشخص می‌شود.

یکی از موضوعات مهم در این مدل، مشخص نمودن مقادیر ضرایب رتبه‌های مربوط به هر بخش می‌باشد. برای این منظور ما از روش تحلیل سلسله مراتبی استفاده کرده‌ایم. برای مقایسه بین گزینه‌های موجود پرسش‌نامه مقایسات زوجی تهیه گشته و در آن سؤالاتی راجع به مقایسه میزان اولویت هر دو زوج از گزینه‌های تجارب، انتشارات و علایق مطرح شده است. مطابق با نظرات ۹ نفر از اساتید دانشگاه بعنوان افراد متخصص، نتایج پرسش‌نامه گردآوری گردیده و بر اساس میانگین هندسی این نتایج روش تحلیل سلسله مراتبی انجام شده است. جدول (۲) نتایج تحلیل سلسله مراتبی و بردار ویژه حاصل را نشان می‌دهد.

جدول (۲) نتایج تحلیل سلسله مراتبی

	تجارب	انتشارات	علایق	بردار ویژه
تجارب	۱	۲/۱۷۲	۳/۰۷	۰/۵۵٤
انتشارات	/٤٦.	۱	۱/۹۳۷	۰/۲۸۳
علایق	۰/۳۲۵	۰/۵۱٦	۱	۰/۱٦۲

ما با تعیین ضرایب a=0.554 برای تجارب، b=0.283 برای انتشارات و c=0.162 برای علایق، رتبه این بخش‌ها را با هم ترکیب نموده و رتبه نهایی یک رزومه را در بازیابی برای نتایج یک پرس‌وجو مشخص کرده‌ایم. نتایج ارزیابی‌ها با استفاده از معیارهای ارزیابی F، MAP، RPREC و Precision-Recall در جداول (۳) و (٤) آمده است.

جدول (۳) نتایج مدل ارائه شده در مقایسه با فضای برداری

مدل	*p@5*	*p@10*	*p@20*	*p@30*	*p@40*
فضای برداری	۸۳/۳	۵۸/۳	۵٤/۱	٤۷/۲	۳۷/۵
مدل ارائه شده	۸۵/۲	۷۳/٦	۵۹/۱	۵۲/٦	٤۳/۳

جدول (٤) نتایج مدل ارائه شده در مقایسه با فضای برداری با استفاده از معیارهای دیگر

مدل	*F*	*MAP*	*R-PREC*
فضای برداری	۲٦/۳	۵٦	۵۹
مدل ارائه شده	۲۷/٤	٦۲/۷	٦٤/۱

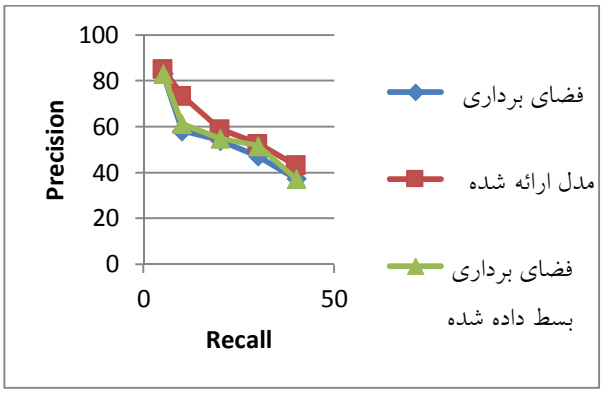
از آنجایی که نتایج مدل فضای برداری همراه با بسط پرس‌و-جو درکل بهتر از مدل فضای برداری مجرد است، به این منظور مدل ارائه شده را با مدل فضای برداری بسط داده شده نیز مورد مقایسه قرار داده‌ایم، تا تفاوت و برتری مدل مشخص گردد. همانطور که در نمودارهای شکل‌های (۵) نمایش داده شده است مدل فضای برداری بسط داده شده نتایج بهتری نسبت به فضای برداری مجرد دارد. نتایج جدول (۵) و (٦) نشان‌دهنده‌ی میزان بهبود در مدل ارائه شده است:

جدول (۵) نتایج مدل ارائه شده در مقایسه با فضای برداری بسط داده شده

مدل	*p@5*	*p@10*	*p@20*	*p@30*	*p@40*
فضای برداری بسط داده شده	*83.3*	*61.4*	*55*	*51.8*	*37.5*
مدل ارائه شده	*85.2*	*73.6*	*59.1*	*52.6*	*43.3*

جدول (٦) نتایج مدل ارائه شده در مقایسه با فضای برداری بسط داده شده با استفاده از معیارهای دیگر

مدل	*F*	*MAP*	*R-PREC*
فضای برداری بسط داده شده	۲٦/۵	۵۷/۸	٦۱/۳
مدل ارائه شده	۲۷/٤	٦۲/۷	٦٤/۱



شکل (۵) مقایسه مدل فضای برداری، فضای برداری بسط داده شده و مدل ارائه شده

این نتایج فرضیه‌ی ما مبنی بر تاثیر متفاوت کلمات موجود در رزومه را اثبات می‌کند. با توجه به این نتایج، درسیستم‌های خبره‌یاب، فروم‌های آنلاین، سیستم‌های پاسخ به سوال، می‌توان با توجه به ماهیت سیستم، در بین داده‌های موجود در پروفایل متخصصین، نوشتجات آنها تمایز قائل شد و با وزن‌دهی به محتوای پر تاثیر، سیستم‌های خبره‌یاب را ارتقاء داد.

۵. نتیجه‌گیری

امروزه سیستم‌های پیشنهادگر نقش مهمی در انتخاب‌های ما دارند. سیستم‌های پیشنهاگر دانش نیز با تصفیه اطلاعات نامرتبط در یافتن پاسخی به نیاز اطلاعاتی کاربران بسیار مفید می‌باشند. فروم‌های آنلاین و سیستم‌های پاسخ به سوال، رشد زیادی در سال‌های اخیر داشته‌اند و مردم نیاز اطلاعاتی خود را با استفاده از اینترنت رفع می‌کنند. سیستم‌های خبره‌یاب سیستم‌هایی هستند که در مواجهه با نیاز اطلاعاتی کاربران، متخصصین دارای دانش را به آنها پیشنهاد می‌کنند. این به نوبه‌ی خود باعث می‌شود که متخصصین با انتقال تجارب و بینش، فهم و دانسته‌های خود را در مورد موضوع منتقل کنند. در این مقاله، مدل جدیدی برای سیستم‌های خبره‌یاب ارائه شده که نتایج بهتری نسبت به مدل فضای برداری به عنوان مدلی مرجع برای سیستم‌های بازیابی اطلاعات دارد و دقت را به نسبت مدل فضای برداری در سیستم پیشنهادگر خبره افزایش داده است. در این مقاله از داده‌های رزومه‌ی اساتید دانشگاه به عنوان مجموعه‌ی داده استفاده شده است. مدل ارائه شده، روشی بر پایه مدل فضای برداری است که بخش‌های مختلف رزومه را به صورت مجزا تحلیل کرده و از ضریب‌های وزنی متفاوت برای بخش‌های مختلف استفاده می‌کند. ایده‌ی استفاده از این روش بر این پایه است که داده‌های موجود در بخشهای متفاوت رزومه دارای اهمیت یکسانی نیستند و برخی داده‌ها در شناسایی تخصص یک فرد دارای اهمیت بالاتری هستند. ما داده‌های سه بخش انتشارات، علایق و تجارب را به ازای تمام رزومه‌ها استخراج کرده و هریک را به صورت مجزا و با مدل فضای برداری تحلیل کرده‌ایم و در نهایت نتایج این سه بخش با هم ترکیب خطی شده‌اند. علاوه بر این ما از ایده‌ی بسط پرس‌وجو استفاده کرده و به هنگام دریافت پرس‌وجوی کاربر با استفاده از ماتریس همبستگی کلمات آن را بسط داده و سپس مراحل جستجو را انجام می‌دهیم.

نتایج حاصل از مدل ارائه شده، نسبت به نتایج مدل فضای برداری مجرد و مدل فضای برداری بسط داده شده، نشان دهنده‌ی بهبود در بازیابی می‌باشد. این نتایج بیانگر این موضوع است که نقش کلمات در رزومه متخصصین متفاوت می‌باشد. به بیان دیگر می‌توان با جدا کردن بخش‌های مختلف رزومه و تعیین میزان اهمیت هر بخش با توجه به نیازهای کاربر، میزان دقت در بازیابی را به مقدار چشمگیری افزایش داد. در این مدل ما با تغییر وزن کلمات موجود در بخش‌های مختلف رزومه، نتایج بازیابی را به مقدار زیادی بهبود داده‌ایم. به عنوان کارهای آینده می‌توان در این زمینه اقدامات زیر را انجام داد: پیشنهاد مدل‌های خبره‌یاب برای زبان فارسی، استفاده از ضرایب مختلف برای ترکیب بخش‌های مختلف، استفاده از روش‌های ترکیب نتایج دیگر مثل فیلترینگ.

مراجع

[1] Lederer A. L., Maupin D. j., Sena M. P., Zhuang Y., *"The technology acceptance model and the world wide web"*, Journal of Decision Support Systems, Elsevier, Vol. 29, No. 3, pp. 269-282, 2000.

[2] Alan W. G., Jian J., Alan S. A., Weiguo F., Zhongju Z., *"ExpertRank: A topic-aware expert finding algorithm for online knowledge communities"*, Journal of Decision Support Systems, Elsevier, Vol. 54, No. 3, pp. 1442-1451, 2013.

[3] Krulwich B., Burkey C., *"The Contactfinder agent: answering bulletin board questions with referrals"*, Proceedings of the 13th National Conference on Artificial Intelligence, Portland, Vol. 1, pp. 10-15, 1996.

[4] Liu X., Wang G. A., Johri A., Zhou M., Fan W., *"Harnessing global expertise: a comparative study of expertise profiling methods for online communities"*, Journal of Information Systems Frontiers, Netherlands: Springer, Vol. 16, No. 4, pp. 715-727, 2012.

[5] Streeter L., Lochbaum K., *"An expert/expert-locating system based on automatic representation of semantic structure"*, Proceedings of the 4th Conference on Artificial Intelligence Applications, San Diego, pp. 345-350, 1988.

[6] Ackerman M. S., Malone T. W., *"Answer Garden: a tool for growing organizational memory"*, Proceedings of the ACM SIGOIS and IEEE CS TCOA Conference on Office Information Systems, Cambridge, Vol. 11, No. 2, pp. 31-39, 1990.

[7] Balog K., Azzopardi L., De Rijke M., *"Formal models for expert finding in enterprise corpora"*, Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, pp. 43-50, 2006.

[8] Balog K., Bogers T., Azzopardi L., De Rijke M., Van den Bosch A., *"Broad expertise retrieval in sparse data environments"*, Proceedings of the 30th Annual International ACM SIGIR Conference, Amsterdam, pp. 551-558, 2007.

[9] Baeza-Yates R., Ribeiro-Neto B., *"Modern Information Retrieval: The Concepts and Technology behind Search"*, 2nd Edition, Boston, USA: ACM Press, 2011.

[10] Cooke R. M., ElSaadany S., Huang X., *"The performance of social network and likelihood-based expert weighting schemes"*, Journal of Reliability Engineering and System Safety, Elsevier, Vol. 93, No. 5, pp. 745–746, 2008.

[11] Manning C. D., Schutze H., *"Foundations of statistical natural language processing"*, Cambridge, USA: MIT Press, 1999.

[12] Mueller Porthmann T., Finke I., *"SELaKT-social network analysis as a method for expert localisation and sustainable knowledge transfer"*, Journal of Universal Computer Science, Springer Verlag, Vol. 10, No. 6, pp. 691–701, 2006.

[13] Lin C., Zhou H., Huang Z., Wang W., *"REC: A novel model to rank experts in communities"*, Proceedings of the 9th International Conference on Web-Age Information Management, Zhangjiajie, pp. 301–308, 2008.

[14] Romney A. K., Weller S. C., Batchelder W. H., *"Culture as consensus: a theory of culture and informant accuracy"*, Journal of American Anthropologist, Wiley-Blackwell, Vol. 88, No. 2, pp. 313-338, 1986.

[15] Zhang J., Ackerman M. S., Adamic L., *"Expertise networks in online communities: structure and algorithms"*, Proceedings of the 16th International World Wide Web Conference (WWW), Banff, pp. 221-230, 2007.

[16] Jurczyk P., Agichtein E., *"Discovering authorities in question answer communities by using link analysis"*, Proceedings of the 16th International Conference on Information and Knowledge Management, Lisboa, Portugal: ACM Press, pp. 919-922, 2007.

[17] Kautz H., Selman B., Shah M., *"Referral web: combining social networks and collaborative filtering"*, Communications of the ACM, Vol. 40, No. 3, pp. 63–65, 1997.

[18] Mattox D., Maybury M., Morey D., *"Enterprise expert and knowledge discovery"*, Proceedings of the 8th International Conference on Human-Computer Interaction, Munich, Vol. 2, pp. 23–27, 1999.

[19] Liu D. R., Chen Y., Kao W., Wang H., *"Integrating expert profile, reputation and link analysis for expert finding in question-answering websites"*, Journal of Information Processing and Management, Elsevier, Vol. 49, No. 1, pp. 312-329, 2013.

[20] Serdyukov P., Rode H., Hiemstra D., *"Modeling relevance propagation for the expert search task"*, The 16th Conference on Text Retrieval (TREC 2007) Enterprise Track, Gaithersburg, pp. 82-87, 2007.

[21] Campbell C. S., Maglio P. P., Cozzi A., Dom B., *"Expertise identification using email communications"*, Proceedings of the 12th International Conference on Information and Knowledge Management, New York, pp. 528-531, 2003.

[22] D'Amore R., *"Expertise community detection"*, Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, pp. 498–499, 2004.

[23] Fu Y., Xiang R., Liu Y., Zhang M., Ma S., *"Finding experts using social network analysis"*, Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, Fremont, pp. 77–80, 2007.

[24] Zhang J., Tang J., Li J., *"Expert Finding in a social network",* Proceedings of the 16th ACM Conference on Information and Knowledge Management, Washington, pp. 1019-1022, 2007.

[25] Saaty, Thomas L., *"How to make a decision: the Analytic Hierarchy Process*", Journal of Operational Research, Berlin: Springer, Vol. 48, No. 1, pp. 9-26, 1990.