دانشکده مهندسی کامپیوتر

طراحی یک سیستم پالایش اطلاعات با استفاده از عاملهای اطلاعاتی تطبیق پذیر

توسط : علی محمدی

پایان‌نامه برای دریافت درجه کارشناسی ارشد

در گرایش هوش مصنوعی.

استادراهنما : دکتر عادل رحمانی

اسفند ۱۳۸۱

تقدیم به

پدر و مادر عزیزم

چکیده

هدف سیستمهای پالایش اطلاعات(سیستمهای IF) پشتیبانی کاربران در یافتن اطلاعات مربوط از یک پایگاه پویای اشیاء اطلاعاتی می‌باشد.سیستمهای IF محاسبات مربوط‌بودن را بر اساس آنچه که پرونده‌کاربر می‌نامند و نمایشهایی از علایق کاربر را نگهداری می‌کند انجام‌می‌دهد. با تغییر علایق کاربر ،در صورت عدم بهنگام‌سازی ، دقت پرونده‌های کاربر کاهش می‌یابد و کیفیت پیش‌بینی هم به مراتب کاهش خواهد یافت. هرچه علایق کاربر پویاتر باشدبهنگام‌سازی سریع پرونده‌های کاربر برای سیستمهای IF جهت کوتاه نمودن مدت‌زمان اُفت کیفیت پیش‌بینی از اهمیت بیشتری برخوردار خواهدبود. هدف این پروژه طراحی یک سیستم پالایش اطلاعات است که بتواند با تغییرات علایق کاربر پرونده کاربر را بهنگام‌سازی نماید. برای رسیدن به این هدف (بهنگام‌سازی مؤثرتر پرونده کاربر) ساختار پرونده کاربر را به گونه‌ای انتخاب می‌کنیم که بتوانیم انواع تغییرات علایق کاربر را در پرونده‌کاربر اعمال نماییم. برای این منظور از مدل «سه توصیفگر» برای پرونده کاربر استفاده می‌کنیم که شامل علایق بلندمدت و کوتاه‌مدت مثبت و کوتاه‌مدت منفی می‌باشد. هر کاربر به زمینه‌های مختلفی علاقه‌مند می‌باشد و باید بتوانیم تنوع علاقه‌مندیهای کاربر را در قالب دسته‌های مختلف علاقه در پروندهٔ کاربر ثبت نماییم.

بعد از انتخاب یک ساختار مناسب برای ثبت علایق کاربر باید به دنبال روشی باشیم که بتوانیم به علایق کاربر پی ببریم.گرفتن بازخورد از کاربر یک روش مناسب برای این منظور می‌باشد. دو نوع بازخورد وجود دارد :

۱- بازخورد صریح (کاربرمستقیماً میزان علاقه‌مندی به یک شیء اطلاعاتی را بصورت یک عدد به سیستم برمی‌گرداند)

۲- بازخورد ضمنی(میزان علاقه‌مندی کاربر به یک شیء اطلاعاتی بدون دخالت مستقیم او وبا مشاهده رفتارش بدست می‌آید). در اینجا از ترکیب دو روش استفاده شده است.

بعداز ایجاد پرونده کاربران باید بتوانیم اشیاء اطلاعاتی را بر اساس این پرونده‌ها کلاس‌بندی نماییم. یعنی میزان تعلق هر شیء اطلاعاتی جدید را به هرکدام از کلاس‌های علاقه‌کاربر تخمین بزنیم. برای این بخش از سیستم از روش بیزین جهت کلاس‌بندی اسناد استفاده نموده‌ایم. بعد از کلاس‌بندی اشیاء ، اشیائی که در یک کلاس علاقه مربوط به یک کاربر تعلق می‌گیرند به کاربر ارائه شده و پس از مشاهده آنها توسط کاربر نتایج کلاس‌بندی از بازخوردهای کاربر حاصل می‌شود که این نتایج در پرونده کاربر تاثیر داده می‌شوند تا دقت کلاس‌بندی‌های بعدی بهتر شود.

با تشکر فراوان از :

- دکتر رحمانی که انجام این پروژه بدون راهنمایی ایشان میسر نبود.
- دکتر کنگاوری و دکتر عبدالله زاده که داوری پایان نامه را بر عهده داشتند.

فهرست مطالب

عنوان صفحه

عنوان صفحه

فصل سوم- مدل کردن کاربر

فصل چهارم-پالایش اطلاعات

عنوان | صفحه

فهرست اشکال

شکل صفحه

فصل اول

بازیابی اطلاعات و پالایش اطلاعات

۱-۱-مقدمه

IR[1] (بازیابی اطلاعات) و IF[2] (پالایش اطلاعات) دو پردازه مجزا هستند که دارای اهداف مشابهی می‌باشند. هردوی آنها با جستجوی اطلاعات سروکار دارند. بر اساس نیازهای اطلاعاتی کاربران که توسط خود آنها به سیستم اعلام می‌شود مجموعه‌ای از اسناد را برمی‌گردانند تا نیازهای اطلاعاتی کاربران برآورده شود. نیاز اطلاعاتی در IR بوسیله پرس‌وجوها و در IF بوسیله پرونده‌های کاربر بیان می‌شوند. مطالعه ما در مورد این پردازه فقط روی اشیائی متمرکز است که اسناد شبه ساختیافته (مانند email) یا بدون ساختار(مانند اشیاء چند رسانه‌ای) می‌باشند. IR قبل از IF می‌باشد زیرا IR قدیمی‌تر می‌باشد و بسیاری از توابع IF براساس IRمی‌باشند. بسیاری از تحقیقاتی که در زمینه IRانجام شده است در IF هم قابل قبول می‌باشند.

۱-۲- مقایسه IR, IF

- سیستمهای IR برای بهبود بعضی نارسائیها در نمایش پرس‌وجوی نیاز اطلاعاتی ایجاد شده‌اند ولی سیستمهای IF فرض می‌کنند که پرونده‌های کاربر دقیق می‌باشند.
- سیستمهای IR معمولاً در یک لحظه توسط یک کاربر استفاده می‌شوند ولی سیستمهای IF مکرراً بوسیله یک کاربر با چند پرونده استفاده می‌گردد.
- سیستمهای IR برای برآورده‌کردن نیازهای اطلاعاتی کوتاه مدت می‌باشند ولی سیستمهای IF برای برآورده کردن نیازهای ثابت یا بلندمدت کاربران می‌باشد.

---

[1] Information Retrieval

[2] Information Filtering

- معمولاً سیستمهای IR روی مجموعه ثابتی از اسناد عمل می‌نمایند ولی سیستمهای IF با یک جریان داده‌ای پویا از اسناد سروکاردارند.
- هدف اصلی IR جمع‌آوری و سازماندهی مجموعه اسنادی می‌باشد که منطبق بر یک نیاز داده شده می‌باشند ولی هدف اصلی IF توزیع اسناد جدید برای کاربرانی می‌باشد که پرونده‌های آنها منطبق بر هم می‌باشد.
- ارائه به‌موقع یک سند در IF دارای اهمیت می‌باشد ولی در IR مهم نمی‌باشد.

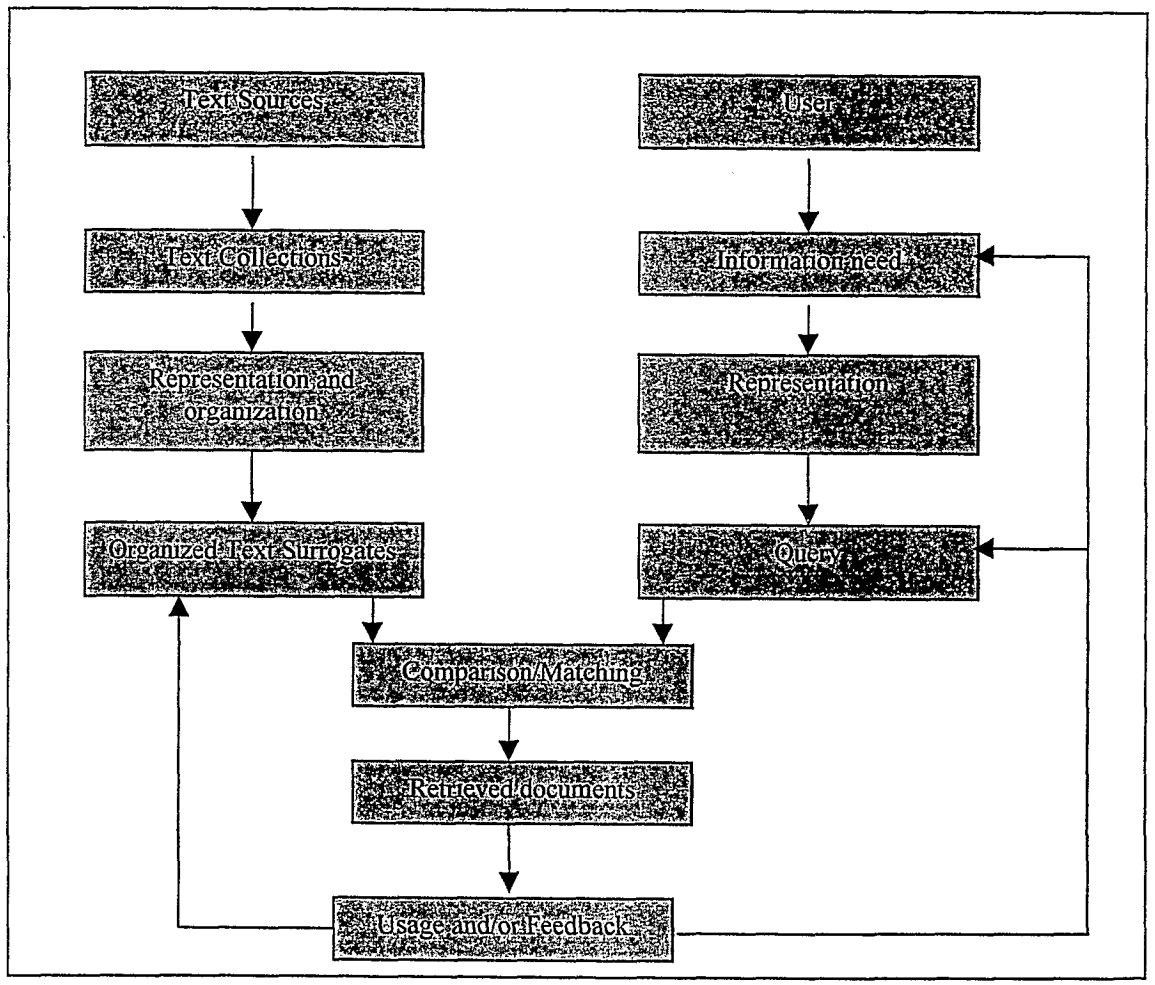
باتوجه به مقایسات فوق مهمترین اختلاف بین IR, IF مربوط به اهداف، کاربرد، کاربران ونوع داده‌ها (ثابت یا پویا) می‌باشد که این دو سیستم روی آن عمل می‌نمایند.

بمنظور فراهم کردن اطلاعات مورد نیاز برای کاربران هردو سیستم IR , IF باید بتوانند نیازهای اطلاعاتی کاربران(هم پرس‌وجو و هم پرونده‌کاربر) و مجموعه اسناد را بروش مناسبی نمایش دهند تا بتوان آنها را باهم مقایسه کرد. بعضی از نمایشهایی که در حال حاضر بکار می‌روند عبارتند از : نمایش‌برداری ، شبکه‌های عصبی ،شبکه‌های معنا و....

مکانیزم مقایسه که برای مطابقت بین نیازهای اطلاعاتی و اسناد بکاربرده می‌شوند بستگی به روشهای نمایش دارد. برای بهبود دقت سیستمهای IR, IFاز مکانیزم بازخورد استفاده می‌کنیم. بعد از اینکه سیستم مجموعه‌ای از اسناد را بعنوان نتیجهٔ جستجو برگرداند کاربر باید برای اسناد انتخاب شده به سیستم بازخورد برگرداند و سیستم نیز براساس این بازخوردها پرس‌وجو یا پرونده کاربر را بهنگام‌سازی نماید تا کارایی سیستم را در جستجوهای بعدی افزایش دهد. بمنظور ارزیابی کارایی سیستم IR, IF از معیارهای Precision , Recall استفاده می‌کنیم. Precision برابر است با نسبت اسنادبازیابی شده که مرتبط با پرس‌وجو یا پرونده کاربر باشد و برای محاسبه آن تعداد اسناد مرتبط برگردانده شده را بر تعداد کل اسناد برگردانده شده تقسیم

می‌کنیم. معیار Recall برابر است با نسبت تعداد اسناد مرتبطبرگردانده‌شده بـه تعـداد کـل اسـناد مرتبطی که در پایگاه دادهٔ منبع وجود دارد. با افزایش تعداد نتایج برگردانـده شـده احتمـال بـالا رفتن Precision هم بیشتر می‌شود ولی در عین حال احتمال وجود نتایج نـامربوط هـم افزایـش می‌یابد.

IR : IR با بازیابی اسنادی که منطبق بر یک پرس‌وجو می‌باشند از یک مجموعه اسناد سروکار دارد. سپس اسناد بازیابی شده رتبه‌بندی شده و به کاربر نشان داده می شوند. در شکل زیر یـک مدل عمومی برای IR نشان داده شده‌است.



شکل۱-۱- یک مدل عمومی برای سیستمهای IR [۲۰]

در این مدل یک کاربر دارای بعضی نیازهای اطلاعاتی یک پرس‌وجـــو را بـه سیســتم IR نشـان می‌دهد. پرس‌وجو نمایشی از نیازهای اطلاعاتی کاربر به زبانی است که توسط سیستم قابل فهم می‌باشد. سپس این پرس‌وجو با اسناد که بوسیله متنهایی جایگزین (لیستی از کلمـات کلیـدی ، عناوین یا چکیده‌ها) شده‌اند مقایسه می‌گردند.

متنی که جایگزین سند شده است را می‌توان بعنوان یک نمـایش سـاختیافته خلاصـه شـده از اطلاعات متنی بدون ساختار در نظر گرفت. مجموعه‌ای از اسناد بعنوان نتیجه مقایســه انتخـاب شده و به کاربر نمایش داده می‌شود. کاربر نیز یا از اسناد استفاده می‌کند و یا اینکه بازخوردهایی به سیستم برمی‌گرداند تا سیستم پرس وجو را بهبود بخشـد. ایـن پـردازه محـاوره تـاوقتی کـه نیازهای کاربر برآورده شوند و یا اینکه کاربر سیستم را ترک کند ادامه می‌یابد.

IF : سیستمهای IF با جریانهای بزرگی از اسناد وارده سروکار دارند که معمـولاً روی منابع دوردست توزیع شده‌اند. IF پرونده‌هایی را که توصیف کننده علایق بلنـد مـدت کـاربران است نگهداری می‌کند. پرونده‌ها ممکن است وصف‌کننده علاقه‌مندیها و غیرعلاقه‌مندیهای کاربر باشد. اسنادی که منطبق بر پرونده کاربر نباشند از مجموعه اسناد جدید حذف می‌شوند. بعنـوان یک نتیجه کاربر فقط آنچه که بعد از عمل حذف در مجموعه اسـناد باقی‌مـانده اسـت را نشـان می‌دهد.