



۱۳۸۲ / ۵ / ۳۰



دانشکده مهندسی کامپیوتر

طراحی یک سیستم پالایش اطلاعات با استفاده از عملیهای اطلاعاتی
تطبيق پذیر

توسط : علی محمدی

پایان نامه برای دریافت درجه کارشناسی ارشد

در گرایش هوش مصنوعی

۴۹۱۸۹

استاد راهنما : دکتر عادل رحمانی

اسفند ۱۳۸۱

تقديم به

پدر و مادر عزیزم

چکیده

هدف سیستمهای پالایش اطلاعات (سیستمهای IF) پشتیبانی کاربران در یافتن اطلاعات مربوط از یک پایگاه پویای اشیاء اطلاعاتی می باشد. سیستمهای IF محاسبات مربوط بودن را بر اساس آنچه که پرونده کاربر می نامند و نمایشهایی از علایق کاربر را نگهداری می کند انجام می دهد. با تغییر علایق کاربر، در صورت عدم بهنگام سازی، دقت پرونده های کاربر کاهش می یابد و کیفیت پیش بینی هم به مراتب کاهش خواهد یافت. هرچه علایق کاربر پویاتر باشد بهنگام سازی سریع پرونده های کاربر برای سیستمهای IF جهت کوتاه نمودن مدت زمان آفت کیفیت پیش بینی از اهمیت بیشتری برخوردار خواهد بود. هدف این پروژه طراحی یک سیستم پالایش اطلاعات است که بتواند با تغییرات علایق کاربر پرونده کاربر را بهنگام سازی نماید. برای رسیدن به این هدف (بهنگام سازی مؤثرتر پرونده کاربر) ساختار پرونده کاربر را به گونه ای انتخاب می کنیم که بتوانیم انواع تغییرات علایق کاربر را در پرونده کاربر اعمال نماییم. برای این منظور از مدل «سه توصیفگر» برای پرونده کاربر استفاده می کنیم که شامل علایق بلندمدت و کوتاه مدت مثبت و کوتاه مدت منفی می باشد. هر کاربر به زمینه های مختلفی علاقه مند می باشد و باید بتوانیم تنوع علاقه مندیهای کاربر را در قالب دسته های مختلف علاقه در پرونده 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 می باشد. دو نوع بازخورد وجود دارد:

۱- بازخورد صریح (کاربر مستقیماً میزان علاقه مندی به یک شیء اطلاعاتی را بصورت یک عدد

به سیستم برمی گرداند)

۲- بازخورد ضمنی (میزان علاقه مندی کاربر به یک شیء اطلاعاتی بدون دخالت مستقیم او و یا

مشاهده رفتارش بدست می آید). در اینجا از ترکیب دو روش استفاده شده است.

بعد از ایجاد پرونده کاربران باید بتوانیم اشیاء اطلاعاتی را بر اساس این پرونده ها کلاس بندی نماییم. یعنی میزان تعلق هر شیء اطلاعاتی جدید را به هر کدام از کلاس های علاقه کاربر تخمین بزنیم. برای این بخش از سیستم از روش بیزین جهت کلاس بندی اسناد استفاده نموده ایم. بعد از کلاس بندی اشیاء، اشیائی که در یک کلاس علاقه مربوط به یک کاربر تعلق می گیرند به کاربر ارائه شده و پس از مشاهده آنها توسط کاربر نتایج کلاس بندی از بازخوردهای کاربر حاصل می شود که این نتایج در پرونده کاربر تاثیر داده می شوند تا دقت کلاس بندی های 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.

صفحه	عنوان
	فصل اول-بازیابی اطلاعات (IR) و پالایش اطلاعات (IF)
۱-۱-۱-۱	مقدمه
۱-۱-۲-۱	مقایسه IF,IR
۱-۳-۱-۳-۱	روشهای بکاررفته برای IR,IF
۱-۳-۱-۳-۱	روشهای سنتی
۱-۳-۱-۳-۱	مطابقت رشته‌ها
۱-۳-۱-۳-۱	مدل بولی
۱-۳-۱-۳-۱	مدل فضای برداری
۱-۳-۱-۳-۱	استفاده از فرهنگلغات
۱-۳-۱-۳-۱	روش فایل امضاء
۱-۳-۱-۳-۱	وارون کردن
۲-۳-۱-۳-۱	روشهای مدرن
۱-۲-۳-۱-۳-۱	روش LSI
۲-۲-۳-۱-۳-۱	روش Connectionist
	فصل دوم - عاملها
۱-۲-۱-۲	مقدمه
۲-۲-۱-۲	عاملهای مستقل
۱-۲-۲-۱-۲	هوش مصنوعی (AI)
۲-۲-۲-۱-۲	شباهتها و تفاوتها بین شیء و عامل
۳-۲-۲-۱-۲	واسطهای انسان و کامپیوتر
۳-۲-۳-۱-۲	عاملهای نرم‌افزاری
۴-۲-۱-۲	عاملهای واسطه
۵-۲-۱-۲	عاملهای اطلاعاتی
۶-۲-۱-۲	عاملهای همکار
۷-۲-۱-۲	عاملهای یادگیرنده در پالایش اطلاعات
۸-۲-۱-۲	عاملهای تطبیق‌پذیر شخصی وب
۱-۸-۲-۱-۲	کلاس‌بند وب
۲-۸-۲-۱-۲	مبدل سند
۳-۸-۲-۱-۲	شبکه عصبی رقابتی

فصل سوم- مدل کردن کاربر

۴۳.....	۱-۳- ابزارها.....
۴۳.....	۱-۳-۱- مقدمه.....
۴۳.....	۱-۳-۲- نمایش دانش.....
۴۶.....	۱-۳-۳- یادگیری ماشین.....
۵۰.....	۲-۳- منابع.....
۵۰.....	۲-۳-۱- مقدمه.....
۵۳.....	۲-۳-۲- شاخص‌های علاقه.....
۵۵.....	۲-۳-۲-۱- شاخص‌های صریح علاقه.....
۵۵.....	۲-۳-۲-۲- شاخص‌های علامت زدن.....
۵۵.....	۲-۳-۲-۳- آزمایشات روی رفتار علامت زدن.....
۵۶.....	۲-۳-۲-۳- شاخص‌های دستکاری کردن.....
۵۷.....	۲-۳-۲-۴- شاخص‌های هدایت.....
۵۷.....	۲-۳-۲-۵- شاخص‌های خارجی.....
۵۸.....	۲-۳-۲-۶- شاخص‌های تکرار.....
۵۸.....	۲-۳-۲-۷- شاخص‌های منفی.....
۵۹.....	۲-۳-۳- تغییر علاقه.....
۶۲.....	۳-۳- خلاصه.....

فصل چهارم- پالایش اطلاعات

۶۴.....	۱-۴- روشها و ابزارهای پالایش اطلاعات.....
۶۴.....	۱-۴-۱- موتورهای جستجو.....
۶۵.....	۱-۴-۲- صفحات خانگی شخصی.....
۶۶.....	۱-۴-۳- ابزارهای مدیریت اطلاعات.....
۶۶.....	۱-۴-۴- مدل عرضه اطلاعات.....
۶۷.....	۱-۴-۵- عاملهای هوشمند.....
۶۸.....	۲-۴- روشهای نمایش اسناد و پرونده کاربر.....
۷۰.....	۲-۴-۱- تکنولوژیهای اصلی برای پالایش.....
۷۰.....	۲-۴-۱-۱- بردارهای کلیدی.....
۷۱.....	۲-۴-۱-۲- n-grams.....
۷۱.....	۲-۴-۱-۳- ساختارهای فوق‌پیوند.....

۷۲.....	۴-۱-۲-۴-پالایش گروهی و مبتنی بر اقتصاد.....	۷۲
۷۲.....	۴-۱-۲-۴-روشهای داده‌کاوی.....	۷۲
۷۳.....	۴-۳-تکنیکهای مدل کردن کاربر.....	۷۳
۷۴.....	۴-۳-۱-پرونده‌سازی بصورت دستی.....	۷۴
۷۶.....	۴-۳-۲-یادگیری با استفاده از مثالهای داده‌شده.....	۷۶
۷۶.....	۴-۳-۳-دسته‌بندی پرونده/کاربر (کلیشه‌کردن).....	۷۶
۷۷.....	۴-۳-۴-یادگیری از طریق مشاهده رفتار کاربر.....	۷۷
۷۸.....	۴-۴-سیستمهای پیشنهاد دهنده مبتنی بر عامل.....	۷۸
۸۰.....	۴-۴-۱-کلاس‌بندی سیستمهای پیشنهاد دهنده اطلاعات.....	۸۰
۸۱.....	۴-۴-۱-۱-پالایش مبتنی بر محتوا.....	۸۱
۸۲.....	۴-۴-۲-۱-پالایش مبتنی بر گروه.....	۸۲
۸۴.....	۴-۴-۳-۱-پالایش مبتنی بر رویداد.....	۸۴
۸۷.....	۴-۴-۴-۱-پالایش مبتنی بر اعتبار.....	۸۷
۸۹.....	۴-۴-۵-۱-پالایش ترکیبی.....	۸۹
۹۰.....	فصل پنجم - نمونه‌هایی از سیستمهای پالایش اطلاعات عامل‌گرا.....	۹۰
	فصل ششم - ارائه یک نمونه سیستم پالایش اطلاعات تطبیق‌پذیر.....	
۱۰۴.....	۶-۱-مقدمه.....	۱۰۴
۱۱۰.....	۶-۲-پرونده‌سازی کاربر.....	۱۱۰
۱۱۰.....	۶-۲-۱-نمایش اسناد.....	۱۱۰
۱۱۲.....	۶-۲-۲-نمایش علایق کاربر.....	۱۱۲
۱۱۶.....	۶-۳-چگونگی تعیین میزان شباهت یک سند با یک مدل از یک رده علاقه.....	۱۱۶
۱۱۶.....	۶-۳-۱-روش یادگیری بیزین.....	۱۱۶
۱۱۶.....	۶-۳-۱-۱-ویژگی‌های یادگیری بیزین.....	۱۱۶
۱۱۸.....	۶-۳-۲-مطابقت روش بیزین با مسئله پالایش اطلاعات.....	۱۱۸
۱۱۹.....	۶-۳-۳-احتمال ساده بیزین.....	۱۱۹
۱۲۲.....	۶-۳-۴-الگوریتم یادگیری.....	۱۲۲
۱۲۳.....	۶-۳-۵-الگوریتم مقایسه.....	۱۲۳
۱۲۴.....	۶-۴-پالایش اسناد.....	۱۲۴
۱۲۶.....	۶-۵-دریافت بازخورد از کاربر.....	۱۲۶
۱۲۶.....	۶-۵-۱-دریافت بازخورد صریح از کاربر.....	۱۲۶
۱۲۸.....	۶-۵-۲-دریافت بازخورد ضمنی از کاربر.....	۱۲۸

۱۳۱.....	۶-۶- یادگیری پرونده کاربر.....
۱۳۱.....	۶-۶-۱- الگوریتم یادگیری.....
۱۳۲.....	۶-۶-۲- بهنگامسازی بردارهای ویژگی توصیفگر.....
۱۳۳.....	۶-۶-۳- بهنگامسازی وزنهای علاقه در مدل کوتاهمدت.....
۱۳۴.....	۶-۶-۴- بهنگامسازی وزنهای علاقه در مدل بلندمدت.....
۱۳۴.....	۶-۶-۵- ایجاد ردههای جدید علاقه.....
۱۳۶.....	۶-۶- خلاصه.....
۱۳۷.....	پیشنهادات.....
۱۳۸.....	مراجع.....

فهرست اشکال

صفحه

شکل

شکل ۱-۱	یک مدل عمومی برای سیستمهای IR	۳
شکل ۲-۱	یک مدل عمومی برای سیستمهای IF	۵
شکل ۱-۲	چرخه اجرایی عامل	۲۴
شکل ۲-۲	عاملهای واسطه	۲۹
شکل ۳-۲	رفتار عامل واسطه (Maes 1994)	۳۳
شکل ۴-۲	بلوک دیاگرام محاوره بین کاربر، عامل و وب	۳۹
شکل ۱-۳	استفاده از نمایش دانش برای محاوره تطبیق پذیر با کاربر	۴۶
شکل ۲-۳	استفاده از یادگیری ماشین برای محاوره تطبیق پذیر با کاربر	۴۸
شکل ۳-۳	دامنه علایق صریح/ضمنی	۵۳
شکل ۴-۳	دسته بندی شاخص های علاقه	۵۴
شکل ۱-۴	پرونده وارد شده بصورت دستی	۷۵
شکل ۲-۴	یادگیری پرونده ها با استفاده از مثالها	۷۵
شکل ۳-۴	کاربران کلیشه ای / دسته بندی شده	۷۷
شکل ۴-۴ الف	مرحله مشاهده از رفتار کاربر	۷۸
شکل ۴-۴ ب	مرحله انطباق رفتار کاربر و محتوای وب	۷۸
شکل ۵-۴	پالایش مبتنی بر محتوا	۸۱
شکل ۶-۴	پالایش مبتنی بر گروه	۸۲
شکل ۱-۶	نمایش سیستم پالایش اطلاعات از دید دستی و کامپیوتری	۱۰۴
شکل ۲-۶	روال کلی پالایش	۱۰۸

بازیابی اطلاعات و پالایش اطلاعات

IR^۱ (بازیابی اطلاعات) و IF^۲ (پالایش اطلاعات) دو پرده مجزا هستند که دارای اهداف مشابهی می‌باشند. هر دو 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 پرده فقط روی اشیائی متمرکز است که اسناد شبه ساخته شده (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 IR می‌باشند. بسیاری از تحقیقاتی که در زمینه IR انجام شده است در IF هم قابل قبول می‌باشند.

۱-۲- مقایسه IR, IF

- سیستمهای IR برای بهبود بعضی نارسائیه‌ها 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 برآورده کردن نیازهای اطلاعاتی کوتاه مدت می‌باشند ولی سیستمهای IF برای برآورده کردن نیازهای ثابت یا بلندمدت کاربران می‌باشد.

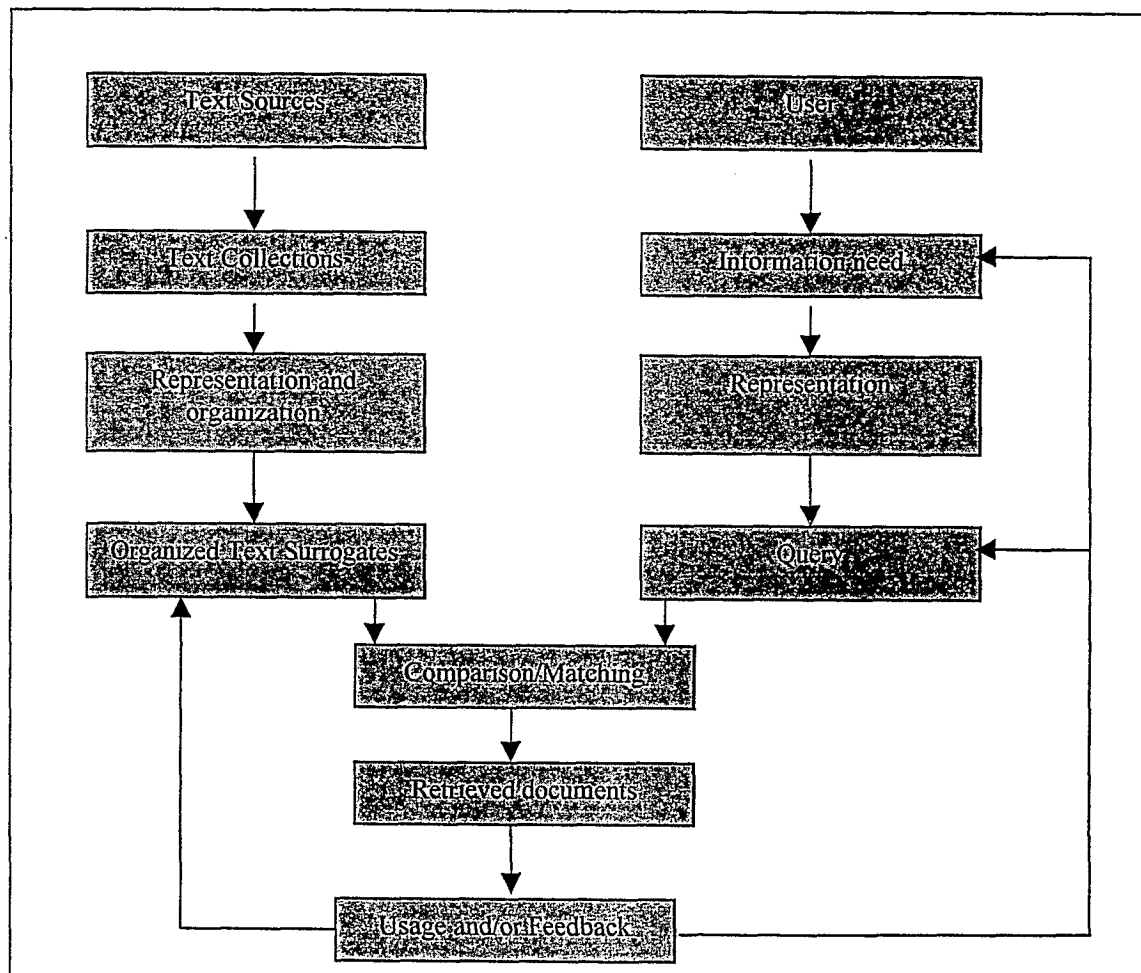
^۱ Information Retrieval

^۲ Information Filtering

- معمولاً سیستمهای IR روی مجموعه ثابتی از اسناد عمل می‌نمایند ولی سیستمهای IF با یک جریان داده‌ای پویا از اسناد سروکار دارند.
 - 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 به موقع 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 و نوع داده‌ها (ثابت یا پویا) می‌باشد که این دو سیستم روی آن عمل می‌نمایند.
- بمنظور فراهم کردن اطلاعات مورد نیاز برای کاربران هر دو 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: نمایش برداری، 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, IF از مکانیزم بازخورد استفاده می‌کنیم.
- بعد از اینکه سیستم مجموعه‌ای از اسناد را بعنوان نتیجه 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 مرتب‌برگردانده شده به تعداد کل اسناد مرتبطی که در پایگاه داده منبع وجود دارد. با افزایش تعداد نتایج برگردانده شده احتمال بالا رفتن Precision هم بیشتر می‌شود ولی در عین حال احتمال وجود نتایج نامربوط هم افزایش می‌یابد.

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 می‌شوند. در شکل زیر یک مدل عمومی برای IR نشان داده شده است.



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

در این مدل یک کاربر دارای بعضی نیازهای اطلاعاتی یک پرس و جو را به سیستم IR نشان می‌دهد. پرس و جو نمایشی از نیازهای اطلاعاتی کاربر به زبانی است که توسط سیستم قابل فهم می‌باشد. سپس این پرس و جو با اسناد که بوسیله متنهایی جایگزین (لیستی از کلمات کلیدی، عناوین یا چکیده‌ها) شده‌اند مقایسه می‌گردند.

متنی که جایگزین سند شده است را می‌توان بعنوان یک نمایش ساختیافته خلاصه شده از اطلاعات متنی بدون ساختار در نظر گرفت. مجموعه‌ای از اسناد بعنوان نتیجه مقایسه انتخاب شده و به کاربر نمایش داده می‌شود. کاربر نیز یا از اسناد استفاده می‌کند و یا اینکه بازخوردهایی به سیستم برمی‌گرداند تا سیستم پرس و جو را بهبود بخشد. این پردازش محاوره تاقیتی که 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 معمولاً روی منابع دوردست توزیع شده‌اند. IF پرونده‌هایی را که توصیف کننده علایق بلند مدت کاربران است نگهداری می‌کند. پرونده‌ها ممکن است وصف کننده علاقه‌مندیها و غیرعلاقه‌مندیهای کاربر باشد. اسنادی که منطبق بر پرونده کاربر نباشند از مجموعه اسناد جدید حذف می‌شوند. بعنوان یک نتیجه کاربر فقط آنچه که بعد از عمل حذف در مجموعه اسناد باقی مانده است را نشان می‌دهد.