به نام خدا

پژوهشکده ارتباطات و فناوری اطلاعات

مطالعه تطبیقی سامانه تحلیل و جستجوی خبر

پروژه: مدیریت طرح

کد پروژه:

مجری: علیرضا یاری

تهیه‌کننده: تیم مدیریت طرح

کد گزارش: P-PD-VAS-SBM-S-011-1.02

تاریخ ارائه: ۹۴/۰۵/۱۷

نسخه/وضعیت:نهایی

شناسنامه گزارش

عنوان: مطالعه تطبیقی سامانه تحلیل و جستجوی خبر		**شماره نسخه:** ۱,۰۲
کد: P-PD-VAS-SBM-S-011-1.02	**نوع گزارش: فنی**	**تاریخ ارائه گزارش:** ۹۴/۰۵/۱۷
نام پروژه: مدیریت طرح	**نوع پروژه:** پژوهشی - کاربردی	
تاریخ شروع: ۹۳/۱۱/۲۰	**تاریخ پایان:** ۹۶/۱۱/۲۰ (۳۶ ماه)	
نام گروه: طرح جویشگر		
کد پروژه: ۹۳۳۲۰۱۲	**شماره و تاریخ قرارداد:** ۹۳/۱۱/۲۰	
مجری: علیرضا یاری	**ناظر/ ناظرین:** : کامبیز بدیع، رامین شکری‌پور و روح‌اله رحمانی	
تهیه کننده/ تهیه کنندگان: تیم مدیریت طرح		

نام و نشانی مجری:

تهران، انتهای خیابان کارگر شمالی، مرکز تحقیقات مخابرات ایران-کد پستی: ۱۴۳۹۹۵۵۴۷۱ - تلفن: ۱۰-۸۸۰۰۵۵۰۸

نام و نشانی حمایت کننده:

تهران، خیابان شریعتی، وزارت ارتباطات و فناوری اطلاعات

ملاحظات: ندارد

چکیده:

تحلیل ۵۰ سایت برتر ایران در سایت الکسا حاکی از این است که حدود ۳۰ درصد از این ۵۰ سایت پرترافیک در ایران سایت‌های خبری هستند. این موضوع بیانگر میزان اهمیت این سایت‌های خبری و استفاده زیاد کاربران ایرانی از این سایت‌ها است. لذا یکی از سرویس‌های جذاب و پرکاربرد برای کاربران که می‌تواند یک مزیت رقابتی برای جویشگر بومی در مقایسه با جویشگرهای محبوبی مانند گوگل باشد، خدمات مبتنی بر اخبار می‌باشد. با توجه به وجود سرویس‌دهنده‌گان خبری متعدد خارجی و داخلی و همچنین حجم زیاد اخبار مربوط به موضوعات خبری مختلف لذا برای پاسخگویی به نیازهای اطلاعاتی کاربران در رابطه با اخبار نیازمند طراحی و پیاده‌سازی سامانه‌های تحلیل و جستجوی خبری می‌باشیم. در این مستند هدف مطالعه و بررسی سامانه‌های تحلیل و جستجوی خبری در سطح کشور و دنیا است که در نهایت به دید مناسبی از وضعیت موجود رسید و با توجه به نقاط ضعف و قوت سامانه‌های موجود در کشور بتوان آن‌ها را توسعه و یا بهبود داد.

کلمات کلیدی: جویشگر بومی

وضعیت گزارش: نهایی	**زبان گزارش: فارسی**
وضعیت دسترسی: عادی	**تعداد صفحات:** ۸۳

چکیده

تحلیل ۵۰ سایت برتر ایران در سایت الکسا حاکی از این است که حدود ۳۰ درصد از این ۵۰ سایت پرترافیک در ایران سایت‌های خبری هستند. این موضوع بیانگر میزان اهمیت این سایت‌های خبری و استفاده زیاد کاربران ایرانی از این سایت‌ها است. لذا یکی از سرویس‌های جذاب و پرکاربرد برای کاربران که می‌تواند یک مزیت رقابتی برای جویشگر بومی در مقایسه با جویشگرهای محبوبی مانند گوگل باشد، خدمات مبتنی بر اخبار می‌باشد. با توجه به وجود سرویس‌دهنده‌گان خبری متعدد خارجی و داخلی و همچنین حجم زیاد اخبار مربوط به موضوعات خبری مختلف لذا برای پاسخگویی به نیازهای اطلاعاتی کاربران در رابطه با اخبار نیازمند طراحی و پیاده‌سازی سامانه‌های تحلیل و جستجوی خبری می‌باشیم. در این مستند هدف مطالعه و بررسی سامانه‌های تحلیل و جستجوی خبری در سطح کشور و دنیا است که در نهایت به دید مناسبی از وضعیت موجود رسید و با توجه به نقاط ضعف و قوت سامانه‌های موجود در کشور بتوان آن‌ها را توسعه و یا بهبود داد.

اطلاعات مرتبط

مستندات مرتبط

شماره مستند	نوع مستند	نام مستند

تغییرات اعمال‌شده در نسخه‌های پیشین

شماره نسخه	تاریخ	تغییرات اعمال شده
۱,۰	۹۴/۰۵/۱۷	آماده‌سازی گزارش
۱,۰۱	۹۴/۰۶/۱۰	اضافه کردن چکیده، کامل کردن مقدمه،اضافه کردن بخش نتیجه‌گیری، اصلاح شکل‌ها و جداول
۱,۰۲	۹۴/۰۶/۲۲	چکیده و نتیجه گیری با توجه به فیدبک ناظرین اصلاح شد

تأییدکنندگان

	نام و نام خانوادگی	تاریخ	امضاء	ملاحظات
مجری پروژه	علیرضا یاری			
تهیه کننده/ تهیه کنندگان	تیم مدیریت طرح			
ناظر پروژه	کامبیز بدیع، رامین شکری‌پور و روح‌اله رحمانی			
مدیر گروه				
مسئول مستندات پژوهشکده	مانا روزی طلب			
رئیس پژوهشکده/ معاون پژوهشی	علیرضا یاری			

اسامی اعضای تیم مدیریت طرح بر اساس حروف الفبا	
۱. محمد آزادنیا	۱۱. مهدی عمادی
۲. محمد مهدی اثنی اشری	۱۲. مژگان فرهودی
۳. شهره جهانبخش	۱۳. محمد مهدی کیخا
۴. مونا داوودی	۱۴. مریم محمودی
۵. غزاله رحمانی فرزین	۱۵. پویان مسعودی فر
۶. فرزانه رحمانی	۱۶. اکبر مقدر
۷. فرزاد زرگری	۱۷. امین میرزائی
۸. محمد صادق زاهدی	۱۸. محمدرضا میرصراف
۹. علی شریفی	۱۹. حمیدرضا نصیری آسایش
۱۰. معصومه عظیم زاده	۲۰. علیرضا یاری
۱۱. طاهره علوی زرگر	

سرفصل مطالب

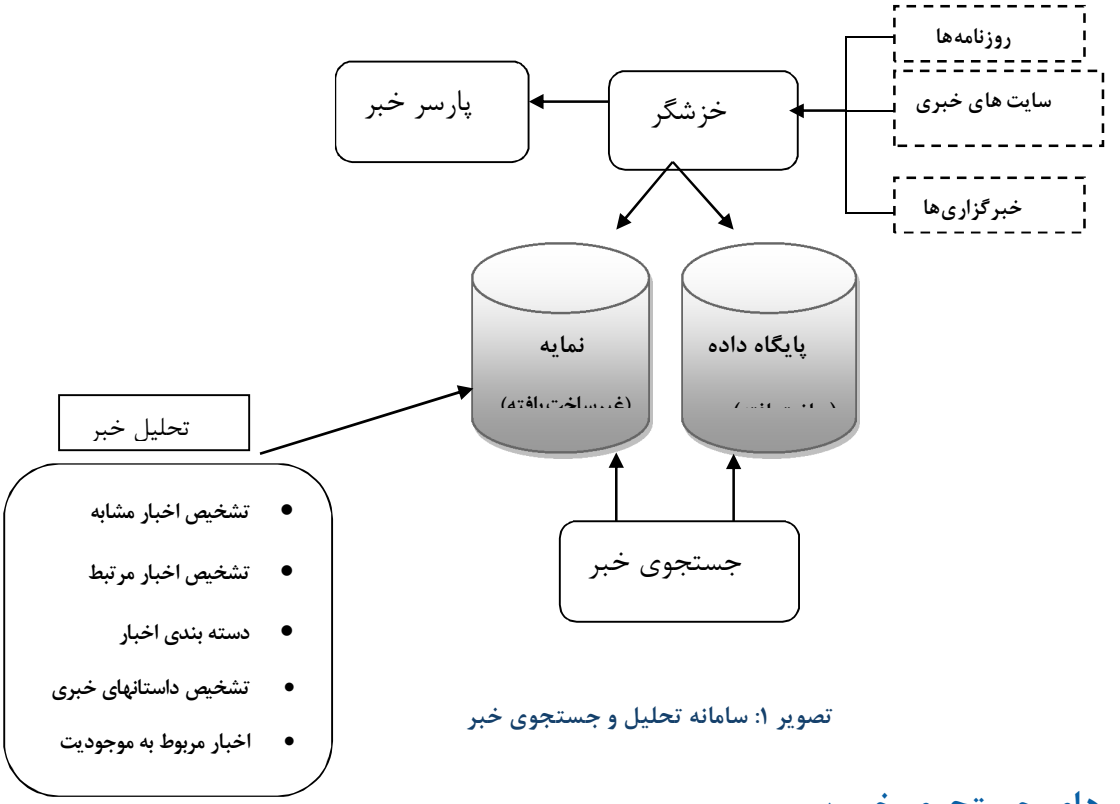
فهرست تصاویر

فهرست جداول

سامانه تحلیل و جستجوی خبر

با توجه به وجود سرویس‌دهنده‌گان خبری متعدد خارجی و داخلی و همچنین حجم زیاد اخبار مربوط به موضوعات خبری مختلف لذا برای پاسخگویی به نیازهای اطلاعاتی کاربران در رابطه با اخبار نیازمند طراحی و پیاده‌سازی سامانه-های تحلیل و جستجوی خبری می‌باشیم. در این مستند به مطالعه و بررسی سامانه‌های تحلیل و جستجوی خبری در سطح کشور و دنیا پرداخته می‌شود. در ابتدا به بررسی معماری کلی سامانه تحلیل و جستجوی خبر پرداخته و بخش‌های مختلف آن معرفی می‌شود. در ادامه به بررسی سایت‌هایی که در کشور خدمات مبتنی بر اخبار ارائه می‌دهند پرداخته می‌شود. ویژگی‌ها و مشخصه‌های اصلی این سایت‌ها مقایسه و تحلیل شده است و بر این اساس نقاط ضعف و قوت این خدمات دهندگان خبری بومی استخراج شده است. علاوه بر سایت‌های خبری بومی، سامانه‌های خبری جهانی نیز بررسی و مقایسه شده است در انتها نیز معیارهای ارزیابی بخش‌های مختلف سامانه تحلیل و جستجوی خبری بررسی شده است.

در شکل ۱ معماری کلی سامانه تحلیل و جستجوی خبری نشان داده شده است. سامانه تحلیل و جستجوی خبر از مولفه‌های خزشگر، پارسر، نمایه و مولفه‌ی تحلیل و جستجوی خبر تشکیل شده است. در مؤلفه خزشگر خبر، صفحات خبری مختلف از سایت‌های خبری، رزونامه‌ها و خبرگزاری‌های معتبر جمع‌آوری می‌شود. سپس مؤلفه پارسر خبر با استفاده از روش‌ها و الگوریتم‌های مختلف پردازش متن، به استخراج بخش‌های مختلف صفحات خبری می‌پردازد. در این مولفه از ابزارهای مختلفی مانند واحدساز، یکسان‌ساز، برچسب‌گذار اجزای کلام، شناسایی کلمات مرکب، ریشه‌یاب و شناسایی جملات فارسی استفاده‌شده و متون برای تحلیل‌های سطح بالاتر آماده می‌شوند. در بخش تحلیل خبر با استفاده از روش‌های مختلف متن کاوی دسته خبر، اخبار مهم، کلمات کلیدی خبر، خبرهای مشابه، وقایع و داستان‌های خبری استخراج می‌شوند. در مؤلفه نمایش خبر نیز اخبار با واسط کاربری مناسب، به کاربرهای مختلف نمایش داده شده و امکان جستجوی اخبار برای کاربران نیز فراهم شده است.



تصویر ۱: سامانه تحلیل و جستجوی خبر

۱-۱ موتورهای جستجوی خبر بومی

در ابتدا لیستی از مهمترین ویژگی‌هایی که یک سامانه تحلیل و جستجوی خبری باید داشته باشد استخراج گردیده و سپس بر اساس این ویژگی‌ها به مقایسه و بررسی سایت‌های خبری موجود در کشور پرداخته شده است. جدول زیر فهرست برخی از معروف‌ترین وب‌سایتهای خبری به همراه ویژگی‌های اصلی آنها را نشان می‌دهد. برای هر ویژگی امتیاز از ۰ تا ۵ در نظر گرفته شده و بر اساس امکانات این سایتها تکمیل شده است. برای تکمیل از داده های شرکتهای که در طرح شرکت کرده‌اند نیز استفاده شده است.

جدول ۱: موتور جستجوی خبر بومی

ردیف	ویژگی	خبر فارسی	شهر خبر	خبر پو	قطره	تی نیوز	خبر خون	خبر پارسی جو	خبر یوز
۱	رتبه در الکسا	۴۹	۲۳	۴۶	۶۴	۷۳	۱۲۳ ۲	۴۷۹	۶۶۳

۲	تشخیص اخبار مشابه(نقل قول ها، اخبار تکراری)	۴,۵	۰	۰	۰	۲	۳	۳	۰
۳	تشخیص اخبار مرتبط(اخبار مرتبط با خبر جاری کاربر)	۴	۰	۰	۰	۰	۴	۴	۳
۴	نمایش لحظه ای اخبار(به صورت زنده)	۲	۴,۵	۳,۵	۰	۴,۵	۵	۲	۱
۵	نمایش پر بیننده ترین اخبار	۴,۵	۵	۴	۰	۴	۴,۵	۳	۲
۶	نمایش پر بحث ترین اخبار	۰	۰	۴	۰	۴	۰	۲	۰
۷	امکان مربوط به کاربران(امکاناتی مانند بحث کردن، دوستی،مشاهده اخبار دوستان و ...)	۰	۰	۰	۰	۰	۰	۰	۰
۸	امکان شخصی‌سازی اخبار مشاهده شده برای کاربر(بر اساس فیلدهای مختلف نوع خبر، مرجع خبر، موضوع خبر ، موجودیت های مختلف مانند افراد، سازمان، مکان ها و ... به صورتی که کاربر اخبار را به صورت فیلتر شده بر اساس پروفایل خود ببیند)	۱,۵	۰	۰,۵	۰	۰	۰	۲	۰
۹	نمایش خبر اصلی (امکانات و نحوه ی نمایش خبر موقع کلیک کاربر بر روی خبر)	۵	۳	۲,۵	۲,۵	۴	۴	۳,۵	۲,۵
۱۰	اخبار چند رسانه ای (تصویر، ویدئو،)	۳	۱,۵	۱	۲	۳,۵	۳,۵	۲	۳,۵
۱۱	واسط کاربری سایت(صفحه ی اول سایت، طراحی سایت و ...)	۴	۲,۵	۲,۵	۱	۳	۵	۳	۴
۱۲	امکانات جستجو (دقت جستجو، جستجوی پیشرفته)	۳,۵	۰	۰	۲	۳,۵	۳,۵	۲	۳,۵
۱۳	دسته بندی اخبار(بر اساس فیلدهای مختلف مانند موضوع، استان، کشور و ...)	۴	۴	۴	۴	۴	۴	۴	۴
۱۴	تشخیص داستانهای خبری (news Stories) و امکانات مربوط به نمایش آن مانند نمایش پربیننده ترین، پر بحث ترین داستان ها، دسته بندی موضوعی داستان های خبری و...	۲,۵	۰	۱	۰	۰	۳	۱,۵	۲
۱۵	اخبار مربوط به موجودیت ها(مانند اشخاص، سازمان و ...)	۰	۰	۰	۰	۱	۰	۰	۲
۱۶	تنوع سایت های خبری پوشش داده شده	۵	۳	۴,۵	۵	۴	۴	۲	۳

۱۷	امکانات تحلیل اخبار شامل مواردی مانند (شناسایی موجددیت ها در اخبار مانند مکان ها ، افراد، سازمان ها و تحلیل های مرتبط با آنها، وتشخیص شبکه انتشار اخبار، کاوش نظرات کاربران در مورد اخبار مربوط به موجودیت های مختلف مانند سازمان یا اشخاص) لطفا در مورد سیستم خبر پیشنهادی خود سایر تحلیل ها در ستون مربوطه آورده شود	۲	۰	۰	۰	۱,۵	۱,۵	۲	۲
		۴۵.۵	۲۳.۵	۲۷.۵	۱۶.۵	۳۹	۳۷.۵	۴۳.۵	۳۲.۵

بر اساس امتیازهای بالا لیست مرتب شده سایتهای خبری به صورت زیر است:

جدول ۲: لیست مرتب شده سایتهای خبری

رتبه	نام موتور جستجو	امتیاز	رتبه الکسا
۱	خبر فارسی	۴۵,۵	۴۹
۲	خبر پارسی جو	۴۳,۵	۴۷۹
۳	تی نیوز	۳۹	۷۳
۴	خبر خون	۳۷,۵	۱۲۳۲
۵	خبر یوز	۳۲,۵	۶۶۳

با بررسی و تحلیل مطالعه صورت گرفته مشخص گردید که ضعف‌های اصلی این سایتها بیشتر مربوط به بخش تحلیل اخبار و شخصی‌سازی اخبار و تشخیص وقایع و داستان های خبری است. در واقع سایت‌های که در کشور خدمات مربوط به خبر ارائه می‌دهند بیشتر در قالب تجمیع‌کننده اخبار هستند و خدماتی مانند تحلیل اخبار و یا شخصی‌سازی اخبار را ارائه نمی‌دهند. با توجه به اهمیت سرویس‌های خبری موتورهای جستجو اصلی موجود در کشور نیز سرویس خبر ارئه می‌دهند. هر یک از این موتورهای جستجو ویژگی‌های منحصربه فرد دارند به عنوان مثال سرویس خبری یوز امکان مشاهده پدیده‌های خبری را برای کاربر فراهم نموده است و یا یک بخش مربوط به کارکاتورهای خبری دارد که ممکن است برای کاربر جذاب باشد همچنین خبرپارسی‌جو نقاط قوتی مانند واسط کاربری مناسب، امکان مشاهده زنده‌ی اخبار، پوشش خبری (جمع‌آوری ودسته‌بندی اخبار مشابه) و نمایش کلیه اخبار متنی و چند رسانه‌ای در یک صفحه، دارد. در کل خبرپارسی‌جو بر اساس معیارهای مشخص شده در جدول ۱ عملکرد بهتری داشته است.

۲-۱ موتورهای جستجوی خبر جهانی

جدول زیر فهرست برخی از معروف‌ترین موتورهای جستجوی خبر به همراه ویژگی‌های اصلی آنها را نشان می دهد

جدول ۳: فهرست معروفترین موتورهای جستجوی خبر

	رتبه جهانی	رتبه محلی	ویژگی‌ها
Google News	۱	۱	نسخه ی اولیه در سال ۲۰۰۲ و نسخه رسمی در سال ۲۰۰۶ شروع شده است ۴۵۰۰ منبع خبری را پوشش داده است نسخه ی محلی برای ۶۰ منطقه و ۲۸ زبان توسعه داده شده است(فارسی ندارد) ۲۵۰۰۰ منتشر کننده خبر برای زبان انگلیسی و ۴۵۰۰ برای سایر زبانها امکان جستجوی خبر و مرتب سازی نتایج بر اساس زمان انتشار خبرها دسته بندی موضوعی خبرها امکان تعریف Alerts برای کلمات کلیدی کاربر و ارسال اخبار مرتبط با ان کلمات بای ایمیل کاربر امکان شخصی سازی اخبار توسط کاربر وجود نسخه ی موبایلی اخبارگوگل امکان دریافت اخبار از طریق فید یا RSS امکان جستجو در آرشیو اخبار مربوط به ۲۰۰ سال گذشته امکان جستجوی پیشرفته بر اساس فیلدهای مختلف مانند خبرگزاری، زمان انتشار خبر، عنوان خبر، بدنه ی خبر ، مکان و ... امکان مشاهده داستان های خبری پربیننده (دسته‌ای از رویدادهای خبری با یکدیگر یک داستان خبری را می‌سازند. مثلاً «بازی‌های آسیائی» یک داستان خبری است. این داستان هر چهارسال یک‌بار تکرار می‌شود و در زمانی که در حال انجام است بیش‌ترین اخبار را به خود اختصاص می‌دهد. اخبار مربوط به این داستان در طول زمان ذخیره‌شده و کاربر می‌تواند تمامی اخبار مربوط به یک داستان را به صورت یکجا مشاهده کند.) نمایش اخبار مرتبط با خبر جستجو شده حذف اخبار تکراری نمایش اخبار بر اساس اخبار مشاهده شده کاربر نمایش اخبار پر بازدید خود خبر اصلی در سایت اصلی منتشر کننده خبر نمایش داده می شود در نتایج گوگل فقط خلاصه خبر نمایش داده می شود علاوه بر ان برای هر عنوان خبری اخبار مربوط به سایتهای مختلف منتشر کننده دسته بندی شده است بر اساس موارد مختلفی مانند : نظر شخصی(Opinion)، تحلیلی(In Depth)، تعداد ارجاع ها (Highly Cited) نمایش داده شده است، کلیه موارد دیگر برای عنوان خبری مانند عکس ها و ویدئوا دسته بندی شده و به صورت مناسبی نمایش داده شده است. نمایش خط زمانی خبر به صورت نمودار
Yahoo! News	۵	۵	رتبه دوم در سایت الکسا در رتب بندی مربوط به سایت های خبری دسته بندی اخبار و امکان نمایش اخبار هر دست متناسب با آن به عنوان مثال دسته ی تکنولوژی به صورت عکس نمایش داده شده است یک عکس مرتبط با خبر که بخشی از خبر نیز در ان نمایان است. و در مان بخش تکنولوژی دسته بندی های دیگری مانند خبرهای مربوط به review، رکت اپل و ... یکی از استراتژی های اصلی یاهو تمرکز بر تولید محتوای اصلی(original content) امکان جستجوی خبر

Reddit	۳۱	۱۰	رتبه اول در سایت الکسا در رتب بندی مربوط به سایت های خبری شبکه اجتماعی خبر است که در آن کاربران ثبت‌نام کرده قادرند اخبار را در قالب لینک یا متن ارسال کنند و آن را با دیگران به اشتراک بگذارند. افراد می توانند گروه تشکیل دهند. یکی از بخش‌های معروف ردیت بخش «من هستم» است که در آن کاربران اعلام می‌کنند که هر چه می‌خواهید از من بپرسید. تاکنون بسیاری از افراد مشهور نظیر، باراک اوباما، جیمی کیمل، ران پال، استیون کلبر، ددماو، زاک براف و نیل استراس در این بخش شرکت کرده‌اند کاربران می‌توانند به پست یا گزارش‌های نوشته شده رأی دهند و جایگاه آن خبر را در صفحات ردیت و صفحه اصلی مشخص کنند. در واقع کاربران هستند که در ردیت اهمیت خبر را تعیین می‌کنند تعداد کارمندان ان ۷۱ نفر است از سال ۲۰۰۵ راه اندازی شده است تعداد بازدیدکننده ردیت در ماه ژولای ۲۰۱۵ حدود ۱۹۵،۲۰۹،۱۶۹ از ۲۱۵ کشور مختلف که تعداد ۸،۰۶۴،۶۷۸،۱۱۲ صفحه را مشاهده کرد اند. امکان جستجوی پیشزفته بر اساس نویسنده خبر، سایت و ... ثبت نام نیازی به ایمیل برای تکمیل ندارد امکان شخصی سازی برای کاربر امکان گذاشتن یک پست خبری بدون رفرنس خارجی تحت عنوان self posts با پایتون توسعه داده شده است

۱-۳ شاخص‌های ارزیابی سامانه تحلیل و جستجوی خبر

هر یک از بخش‌های سامانه تحلیل و جستجوی خبر را می‌توان به صورت مجزا بر اساس معیارهای مختلف ارزیابی نمود. در ادامه برخی از معیارها و آزمون‌های ارزیابی بخش‌های مختلف توضیح داده خواهد شد.

۱-۳-۱ معیارهای ارزیابی خزشگر سامانه تحلیل و جستجوی خبری

۱- سرعت: تعداد و حجم خبرهای جمع‌آوری شده در واحد زمان

۲- پوشش خبر و تازه بودن: در این معیار تعداد خبرگزاری‌ها، روزنامه‌ها و سایتهای خبری پوشش داده شده مورد بررسی قرارمی گیرد و یک مدت زمان ماکزیمم برای فاصله انتشار اگهی و جمع‌آوری آن توسط خزشگر تخمین زده می‌شود. بدیهی است هر چقدر زمان تاخیر کمتر باشد کارایی خزشگر بیشتر است.

۳- تحمل‌پذیری خطا: در این بخش از ارزیابی، رفتار خزشگر درمواجه با انواع خطاهای شبکه، دام‌ها حلقه بررسی می شود.

۱-۳-۲ معیارهای ارزیابی بخش پارسر و استخراج اطلاعات

مهمترین معیارها برای سنجش کارایی این ماژول معیارهای دقت(precision)، بازآوری (Recall) و معیار F است. که به صورت زیر تعریف می‌شوند:

$$\text{Precision} = \frac{NNN\ NNNN\ NN\ NNNNNNNN\ NNNNNNNNNN\ NNNN}{NNN\ NNN\ NN\ NNNNNNNN\ NNNNNN\ NNNN}$$

$$NNNNN = \frac{NNN\ NNNN\ NN\ NNNNNNNN\ NNNNNNNNNN\ NNNN}{NNN\ NNN\ NN\ NNNNNN\ NNNN}$$

معیار F که میانگین هارمونی دقت وبازآوری است نیز به صورت زیر تعریف می شود:

$$N = \frac{۲\text{PR}}{\text{P} + \text{R}}$$

۱-۳-۳ معیارهای ارزیابی بخش جستجوی خبر

معیار دقت در مرتبه n یا P@N: این معیار کسری از خبرهای مرتبط به پرس‌وجو را که در n رتبه اول بازیابی شده‌اند نشان می‌دهد. مقادیر مرسوم برای n عبارتند از :۵، ۱۰، ۲۰.

معیار متوسط میانگین دقت(MAP): میانگین دقت(average precision) در مرتبه n را برای مقادیر مختلف n برای یک پرس وجو داده شده نشان می دهد. متوسط میانگین دقت را برای یک مجموعه پرس وجو (حداقل ۵۰ پرس وجو) نان می دهد.

$$\text{MAP}(\text{Q}) = \frac{۱}{|\text{Q}|} \sum_{NN۱}^{|N|} \frac{۱}{\text{m}_N} \sum_{NN۱}^{N_N} N \ \text{Precision}(\text{R}_N)$$

معیار NDCG: در این معیار، علاوه بر اینکه مرتبط بودن هر خبر مشخص می‌شود، رتبه آن نیز در عمل مقایسه، تاثیرگذار است؛ از آنجایی که کاربران علاقمند به رؤیت صفحات مرتبط در بالای لیست هستند، در این روش نتایجی که ارتباط بیشتری با پرس‌وجوی کاربر دارند در صورتی که رتبه بالاتری داشته باشند از اهمیت بیشتری در محاسبه برخوردارند. در محاسبه NDCG ابتدا باید مقدار CGرا محاسبه کنیم برای این منظور مجموع میزان ارتباط هر نتیجه ارائه شده در لیست نتایج با پرس‌وجوی کاربر را به دست می‌آوریم.

$$\text{CG}_N = N \sum_{NN۱}^{N} \text{rel}_N$$

که در اینجا rel_i میزان ارتباط نتیجه iام با پرس‌وجو می‌باشد. عددی که به کمک این تابع به دست می‌آید رتبه‌بندی سیستم را مد نظر قرار نمی‌دهد و تنها معیاری برای مرتبط بودن هر سند است، بنابراین اگر نتیجه‌ای که ارتباط کمتری با پرس‌وجو دارد در رتبه بالاتری نسبت به نتیجه مرتبط‌ترقرار بگیرد در مقدارCG تفاوتی ایجاد نمی‌شود. با توجه به این مشکل از DCG برای لحاظ نمودن رتبه‌بندی استفاده می‌کنیم.

$$\text{DCG}_N = \text{rel}_۱ + N \sum_{NN۲}^{N} \frac{\text{rel}_N}{\log_۲ \text{i}}$$

به این ترتیب اگر نتایجی که ارتباط بیشتری دارند در رتبه پایین‌تر قرار بگیرند سودمندی و کارایی سیستم کاهش می‌یابد. با توجه به اینکه اندازه نتایج جستجو برای پرس‌وجوهای مختلف متفاوت است بنابراین برای مقایسه کارایی موتورهایی جستجو باید DCG را نرمال کنیم.

$$\mathrm{nDCG}_N = \frac{\mathrm{DCG}_N}{\mathrm{IDCG}_N}$$

در اینجا IDCGp همان DCG ایده‌آل در نقطه p است.

۱-۳-۴ معیارهای ارزیابی بخش تحلیل خبر

هر یک از بخشهای تحلیل خبر ارزیابی خاص خودش را نیاز دارد. در ادامه معیارهای مربوط به برخی از این بخش‌ها آورده می شود:

بخش خوشه بندی اخبار: روش‌های ارزیابی خوشه‌های حاصل از خوشه‌بندی را به صورت سه دسته تقسیم می‌کنند که عبارتند از:

- معیارهای خارجی (External Criteria)
- معیارهای درونی (Internal Criteria)
- معیارهای نسبی (Relative Criteria)

هم معیارهای خارجی و هم معیارهای درونی بر مبنای روش‌های آماری عمل می‌کنند و پیچیدگی محاسباتی بالایی را نیز دارا هستند. معیارهای خارجی عمل ارزیابی خوشه‌ها را با استفاده از بینش خاص کاربران انجام می‌دهند. معیارهای درونی عمل ارزیابی خوشه‌ها را با استفاده از مقادیری که از خوشه‌ها و نمای آنها محاسبه می‌شود، انجام می‌دهند.

پایه معیارهای نسبی، مقایسهٔ بین شماهای خوشه‌بندی (الگوریم به علاوهٔ پارامترهای آن) مختلف است. یک و یا چندین روش مختلف خوشه‌بندی چندین بار با پارامترهای مختلف روی یک مجموعهٔ داده اجرا می‌شوند و بهترین شمای خوشه‌بندی از بین تمام شماها انتخاب می‌شود. در این روش مبنای مقایسه، شاخص‌های اعتبارسنجی (Validity-Index) هستند.

شاخص‌های اعتبارسنجی برای سنجش میزان صحت (Goodness) نتایج خوشه‌بندی به منظور مقایسهٔ بین روشهای خوشه‌بندی مختلف یا مقایسهٔ نتایج حاصل از یک روش با پارامترهای مختلف مورد استفاده قرار می‌گیرند مانند:

- Davies-Bouldin
- Calinski-Harabasz
- Dunn index
- R-square index
- Hubert-Levin(C-index)

- Krzanowski-Lai
- Hartigan index
- Root-mean-square standard deviation (RMSSTD) index
- Semi-partial R-squared (SPR) index
- Distance between two clusters (CD) index
- weighted inter-intra index
- Homogeneity index
- Separation index

بخش خلاصه‌سازی اخبار:

Precisionو recall به عنوان معیارهای متداول در خلاصهسازی مطرح شده است recall .برابر است با تقسیم تعداد جملاتی که توسط سیستم درست تشخیص داده شده است بر تعداد جملاتی که توسط انسان درست تشخیص داده شده اند Precision .برابر است با تقسیم تعداد جملاتی که توسط سیستم درست تشخیص داده شده‌اند بر تعداد کل جملاتی که توسط سیستم برای خلاصه ایجاد شده‌اند.

$$Precision = \frac{SentenceCount(Summary_{SystemExtracted} \cap Summary_{Ideal})}{SentenceCount(Summary_{SystemExtracted})}$$

$$Recall = \frac{SentenceCount(Summary_{SystemExtracted} \cap Summary_{Ideal})}{SentenceCount(Summary_{ideal})}$$

مشخص است که هر چقدر طول خلاصه افزایش یابد، تفاوت بین متن خلاصه شده توسط انسان و متن خلاصه شده اتوماتیک افزایش می یابد. در حالت کلی ارزیابی به دو گروه اصلی و فرعی تقسیم می‌شود. در ارزیابی اصلی، کیفیت خلاصه سازی از طریق مقایسه آن با خلاصه های ایده آل دستی انجام می شود. در ارزیابی فرعی مشخص می شود خلاصه ایجاد شده چقدر روی انجام کارهای دیگر تاثیر گذار بوده است. از جمله این کارها میتوان به پرسش‌وپاسخ، درک مطلب و مرتبط بودن سند به موضوع خاص نام برد. در صورتی که خلاصه ایجاد شده از نوع چکیده باشد ارزیابی از طریق مقایسه محتوایی با خلاصه دست ساز انسان می باشد(ارزیابی محتوایی). این نوع ارزیابی نیاز به تحلیل زبان شناختی دارد. مثلا فاصله بین بردار تکرار کلمات خلاصه ایجاد شده توسط انسان و سیستم محاسبه می شود. نوع دیگر ارزیابی، ارزیابی موضوعی می باشد. در این حالت از افراد خواسته میشود که خلاصه سیستم را از دو دیدگاه ارزیابی و امتیاز دهی نمایند. اول اینکه خلاصه سیستم به چه میزان از محتویات مهم مقاله اصلی را می پوشاند و دوم آنکه خلاصه سیستم تا چه انداز خوانا می‌باشد.

شناسایی موجودیت ها:

معیارهای دقت(precision)، بازآوری (Recall) و معیار F مهمترین معیارهای به صورت کلی برای سیستم های استخراج اطلاعات هستند. که در این بخش قابل استفاده هستند.

شناسایی وقایع و داستان های خبری:

موارد مختلفی در پژوهش‌های قبلی انجام شده است مانند موارد زیر.

۲,۱ New Event Detection (NED) (What's new):

۲,۲ Event Tracking

۲,۳ Event Summarization (What happened)

در اکثر این پژوهش‌ها از معیارهای دقت(precision)، بازآوری (Recall) استفاده شده است که در هر بخش متناسب با آن تعریف می شود.

۱-۳-۵ معیارهای ارزیابی نیازهای غیرکارکردی

- دسترس‌پذیری[۱]
  - میزان پاسخگویی سیستم در قبال پرس‌وجوهای ارسال‌شده
  - نداشتن شکست در سیستم
- کارایی[۲]
  - زمان پاسخگویی به پرس‌وجوها
  - توانایی در پاسخگویی به کاربر همزمان
- آزمون گرافیک واسط کاربری[۳]
- قابلیت همکاری[۴]
  - داشتن وب‌سرویس مناسب

۱-۴ نتیجه‌گیری

با توجه اهمیت سرویس تحلیل و جستجوی خبر تمامی موتورهای جستجو در کشورهای مختلف سرویس خبر ارائه داده‌اند. تحلیل سایت‌های پرترافیک کشور نیز حاکی از این است که سایت‌های خبری جزو سایت‌های پربازدید در کشور است. لذا با توجه به تعداد زیاد خبرگزاری‌ها و سایتهای خبری و همچنین انتشار حجم عظیمی از انواع اخبار از دسته‌های موضوعی مختلف برای پاسخگویی به نیازهای اطلاعاتی کاربران نیازمند طراحی راه‌کارهای مناسب برای تسهیل کاربران در پیداکردن اطلاعات مورد نظر خود خواهیم بود. بر رسی سایت‌های خبری موجود در کشور نشان می‌دهد که اکثر این سایت‌ها به صورت تجمیع‌کننده اخبار هستند و خدماتی مانند تحلیل و جستجوی اخبار و یا امکان شخصی‌سازی اخبار برای کاربران فراهم نشده است. لذا نیازمند توسعه‌ی سامانه تحلیل و جستجوی خبر خواهیم بود البته هدف از این سامانه توسعه‌ی خدماتی است که با پوشش سایت‌های خبری و خبرگزاری‌ها و

---

[۱] Availability

[۲] Performance

[۳] Gui Testing

[۴] Interoperability

روزنامه‌های مختلف امکان انواع جســـتجوی خبر را برای کاربر فراهم کرده و با تحلیل خودکار اخبار کاربر را در رفع نیازهای اطلاعاتی خود کمک نماید. در این سامانه هدف فقط ایجاد یک تجیمع کننده اخبار نیست که اخبار را از تمام ســایتهای خبری اســتخراج کرده و به کاربر نمایش دهد بلکه ســامانه‌ی طراحی شــده باید کاربر را در یافتن اخبار مورد نظر خود کمک کرده و با امکاناتی که ارائه می نماید رفع نیازهای اطلاعاتی خبری کاربر را تســـهیل نماید. به عنوان مثال با تحلیل خودکار محتوای اخبار از خبرگزارهای مختلف اخبار تکراری را حذف کرده همچنین خبرهای مرتبط با پرس وجوی کاربر را نمایش دهد و یا امکان تحلیل نظرات کاربران در مورد اخبار را داشته باشد .