
بسم الله الرحمن الرحيم

دانشکده مهندسی کامپیوتر و فناوری اطلاعات

گروه هوش مصنوعی

سیستم مکانیزه طبقه‌بندی اخبار در بستر وب

دانشجو: ملیکا یعقوبی

استاد راهنما:
جناب آقای دکتر حمید حسن پور

استاد مشاور:
جناب آقای دکتر مرتضی زاهدی

پایان نامه ارشد جهت اخذ درجه کارشناسی ارشد

زمستان 1391

دانشگاه صنعتی شاهرود

دانشکده مهندسی کامپیوتر و فناوری اطلاعات
گروه هوش مصنوعی

پایان نامه کارشناسی ارشد خانم ملیکا یعقوبی
تحت عنوان: سیستم مکانیزه طبقه‌بندی اخبار در بستر وب

در تاریخ ..... توسط کمیته تخصصی زیر جهت اخذ مدرک کارشناسی ارشد مورد ارزیابی و با درجه ..... مورد پذیرش قرار گرفت.

اساتید راهنما	امضاء	اساتید مشاور	امضاء
دکتر حمید حسن‌پور		دکتر مرتضی زاهدی	

اساتید داور	امضاء	نماینده تحصیلات تکمیلی	امضاء
دکتر علی اکبر پویان		مهندس علی بازقندی	
دکتر علیرضا الفی			

تقدیم به پدر و مادرم،

که از نگاهشان صلابت

از رفتارشان محبت

و از صبرشان ایستادگی را آموختم.

تشکر و قدردانی

سپاسگزار کسانی ام که سرآغاز تولدم بودند،

از یکی زاده شدم و از دیگری جاودانه.

پدر و مادری که تار مویی از آنها به پای من سیاه نماند،

استادی که سپیدی علم و اندیشه را بر سیاهی جهل در زندگیم نگاشت.

با سپاس فراوان از استاد راهنمای بزرگوارم، جناب آقای دکتر حسن‌پور، به پاس رهنمودها، کلام و حضورشان، که گذران سخت‌ترین روزها را برایم ممکن ساخت. و قدردانی از استاد مشاور ارجمند، جناب آقای دکتر زاهدی که با نظرات سازنده و اندرزهای حکیمانه‌شان، مرا راهنمایی نمودند.

همچنین بر خود لازم می‌دانم از اعضای خانواده و تمامی دوستانی که در اتمام این پروژه به من یاری رساندند، کمال سپاس را به جای آورم.

تعهد نامه

اینجانب **ملیکا یعقوبی** دانشجوی دوره کارشناسی ارشد رشته **هوش مصنوعی** دانشکده کامپیوتر و فناوری اطلاعات دانشگاه صنعتی شاهرود نویسنده پایان نامه **سیستم مکانیزه طبقه‌بندی اخبار در بستر وب** تحت راهنمائی **جناب آقای دکتر حمید حسن‌پور** متعهد می‌شوم .

- تحقیقات در این پایان نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است .
- در استفاده از نتایج پژوهشهای محققان دیگر به مرجع مورد استفاده استناد شده است .
- مطالب مندرج در پایان نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است .
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می باشد و مقالات مستخرج با نام « دانشگاه صنعتی شاهرود » و یا « Shahrood University of Technology » به چاپ خواهد رسید .
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان نامه تأثیرگذار بوده اند در مقالات مستخرج از پایان نامه رعایت می گردد.
- در کلیه مراحل انجام این پایان نامه ، در مواردی که از موجود زنده (یا بافتهای آنها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است .
- در کلیه مراحل انجام این پایان نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری ، ضوابط و اصول اخلاق انسانی رعایت شده است .

تاریخ: 1391/11/15

امضای دانشجو: **ملیکا یعقوبی**

مالکیت نتایج و حق نشر

- کلیه حقوق معنوی این اثر و محصولات آن (مقالات مستخرج ، کتاب ، برنامه های رایانه ای ، نرم افزار ها و تجهیزات ساخته شده است) متعلق به دانشگاه صنعتی شاهرود می باشد . این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود .
- استفاده از اطلاعات و نتایج موجود در پایان نامه بدون ذکر مرجع مجاز نمی باشد،

چکیده

رشد روزافزون و گسترش چشمگیر تعداد وب‌سایت‌ها و حجم بالای داده‌های موجود در اینترنت، یکی از چالش‌هایی است که دهکده جهانی وب با وجود مزایای بیشمارش برای انسان به همراه آورده است. از طرف دیگر، امروزه نیاز و تمایل برای کسب دقیق و به موقع اطلاعات پیرامون حوادث جاری، به منزله برنامه‌ریزی جهت گذران زندگی، چه در سطح شخصی و چه در سطح سازمانی و سیاسی از اهمیت بسیار بالایی برخوردار است.

سرعت بالای گسترش اخبار در اینترنت و عدم امکان کنترل آنها پس از انتشار، افزایش چشمگیر منابع خبری و حجم بالای اخبار منتشر شده در موضوعات و فیلدهای مختلف، امکان پیگیری مداوم اخبار به صورت دستی را غیر ممکن نموده است. این حجم بالا نه تنها در بدست آوردن اطلاعات مورد نیاز به کاربران کمک نمی‌کند، بلکه باعث سردرگمی و ابهام بیشتر آن‌ها نیز می‌گردد، تا آنجا که نیاز به سیستمی برای مدیریت و در نهایت متمایز ساختن اسناد خبری از یکدیگر و دسته‌بندی آن‌ها در گروه‌های متشابه در این بستر به چشم می‌خورد.

در این پروژه برای رفع این معضل، از تلفیق روش‌های وب‌کاوی و دسته‌بندی داده‌های متنی برای مدیریت بهینه اسناد خبری بهره گرفته‌شده است. قسمتی از متن اسناد خبری در ابتدا به کمک یک خدمتگزار وب از سایت‌های خبری مورد تایید و مشخص برداشته شده و پس از آموزش سیستم، پیش‌پردازش‌های مورد نیاز بر روی اسناد قابل تست صورت گرفته، هر سند خبری به دسته‌های مربوطه و انتخابی کاربر ارجاع داده شده و نتیجه به صورت لیستی از اخبار دسته‌بندی شده نمایش داده می‌شود. از دو عامل پویایی اسناد خبری و ناهمگونی‌های موجود در زبان فارسی می‌توان به عنوان اساسی‌ترین چالش‌های موجود در روند کار نام برد.

تا کنون بیشترین تمرکز در مبحث دسته‌بندی اسناد بر روی استفاده از معیارهای شباهت متفاوت و مقایسه عملکرد آنها بر روی ویژگی‌های انتخابی بوده است. همچنین در اکثر موارد از فرکانس تکرار لغات در متن و ارتباط آنها با مجموعه اسناد تحت آزمایش و در چند مورد نیز از تعداد اسناد درون

گروهی به عنوان ویژگی انحصاری متن استفاده شده است. در حالی که در این پروژه، بیشترین تاکید بر روی آماده‌سازی و نرمال‌سازی داده‌های قابل پردازش، ادغام دیکشنری‌های کمکی به منزله افزایش اهمیت کلمات کلیدی در گروه‌ها و در نهایت توجه به فرکانس تکرار لغات هر سند -به صورت مستقل از دیگر اسناد- در گروه‌های مختلف صورت گرفته شده است. همچنین به منزله انتساب سند حد آستانه‌ای برای تعیین حداقل میزان شباهت در نظر گرفته شده ، که تعلق سند به بیش از یک گروه را ممکن می‌سازد. نتیجه بدست آمده حاکی از موفقیت روش پیشنهادی بر روی داده‌های خبری موجود در وب می‌باشد.

کلمات کلیدی: دسته بندی، ریزاسناد خبری ، متن کاوی، وب کاوی

فهرست مطالب

فصل اول: مقدمه		1
1-1	پیشگفتار و ضرورت پروژه	2
2-1	تعریف پروژه	3
3-1	ساختار پروژه	6
فصل دوم: مبانی نظری و پایه		7
1-2	مقدمه	8
2-2	دنیای وب	8
1-2-2	وب کاوی	9
2-2-2	روش های دسترسی به اطلاعات وب	13
3-2-2	کدهای وضعیت HTTP و خطاهای سرور	24
3-2	دسته بندی اسناد	26
1-3-2	خوشه‌بندی	26
2-3-2	دسته‌بندی	34
3-3-2	دسته بندی اسناد متنی	36
1-3-3-2	آماده سازی و پیش پردازش داده‌ها	37
2-3-3-2	وزن دهی به کلمات و ایجاد ماتریس	39
3-3-3-2	کاهش داده و کاهش بعد	44
4-3-3-2	معیارهای شباهت	44
4-2	دسته بندی اسناد وب	46
1-4-2	مراحل خوشه بندی نتایج وب	49
2-4-2	معرفی موتورهای دسته بند	54
3-4-2	سیستم دسته بندی اسناد در مقابل موتورهای دسته بند وب	59

4-4-2	ارزیابی عملکرد موتورهای دسته بندی	61
5-2	دسته بندی اسناد خبری در وب	63
1-5-2	تفاوت اسناد وب و اسناد خبری	64
2-5-2	چالش های موجود در وب	65
3-5-2	بررسی دیدگاه های متفاوت	65
فصل سوم: مروری بر تحقیقات		68
1-3	مقدمه	69
فصل چهارم: روش تحقیق		78
1-4	مقدمه	79
2-4	اخبار موجود در وب	80
3-4	پویش صفحات و جستجو لینک‌های خبری	82
1-3-4	مدیریت خطاهای احتمالی به کمک کدهای وضعیت HTTP	84
2-3-4	انتخاب قطعه خبری	86
4-4	فاز آموزش	88
1-4-4	جمع آوری اسناد آموزشی	88
2-4-4	تطبیق و تعیین گروه‌ها	90
3-4-4	آماده سازی اسناد آموزشی	91
4-4-4	ایجاد پایگاه لغات	95
1-4-4-4	ادغام پایگاه های لغت کمکی	95
2-4-4-4	تعیین و حذف کلمات پر کاربرد	96
3-4-4-4	تعلق کلمات به گروه‌ها	97
5-4	فاز تست	98
1-5-4	تهیه قطعات	98

2-5-4	آماده‌سازی اسناد آزمایشی	99
3-5-4	وزندهی به کلمات	99
4-5-4	انتساب اسناد و تعیین مقدار آستانه	100
5-5-4	بازسازی مجدد پایگاه لغات	104
6-4	نمای کلی برنامه	104
7-4	ارزیابی	105
فصل پنجم: نتیجه گیری و پیشنهادات		109
1-5	مقدمه	110
2-5	تحلیل نتایج	111
3-5	اقدامات آینده	114
مراجع		116

فهرست اشکال

شکل(2-1) تغییر ساختار ارائه اسناد وب از حالت لیست رتبه‌بندی شده به دسته‌های مجزا 23

شکل(2-2) ماتریس سند-لغت، نمایانگر وزن هر لغت در سند متناظر 40

شکل(2-3) مراحل خوشه بندی نتایج وب 50

شکل(2-4) مقایسه میانگین زمان اجرا و دقت برای تعداد دسته و درخواست مشابه در کل متن و قطعات وب 52

شکل(2-5) میانگین دقت الگوریتم های دسته بندی برای 10 مجموعه سند 53

شکل(2-6) مقایسه روش های STC , Lingo 54

شکل(2-7) تصویری از خروجی موتور دسته بند carrot2 در مورد درخواست Iran 58

شکل(2-8) ایجاد فضایی برای تشخیص گروه خبر توسط موتور دسته بند نتایج 66

شکل(2-9) دسته بندی موضوعات خبری توسط موتور دسته بند 66

شکل(2-10) ایجاد جریان خبری توسط موتور دسته بند نتایج 67

شکل(4-1) نمای کلی برنامه 105

شکل(5-1) نمونه ای از تناقض در گروه بندی پیشنهادی خبرگزاری‌ها 112

فهرست جداول

جدول(2-1)	نحوهٔ تشخیص عملیات مناسب برای حل مشکلات موتورهای جستجو	12
جدول(2-2)	ابزارها و روش‌های وب کاوی	13
جد.ل(2-3)	خلاصه ای از روش‌های مورد استفاده در خوشه‌بندی اسناد	28
جدول(2-4)	روش‌های وزندهی بر اساس فرکانس تکرار لغت در سند	41
جدول(2-5)	روش‌های وزندهی بر اساس معکوس فرکانس تکرار لغت در اسناد و ترکیب آن با روش محلی	42
جدول(2-6)	روش‌های وزندهی بر اساس فرکانس تکرار لغت در دسته	43
جدول(2-7)	روش‌های اندازه‌گیری معیار شباهت بر مبنای فاصله	45
جدول(2-8)	روش‌های اندازه‌گیری معیار شباهت بر مبنای نزدیکی	46
جدول(2-9)	خلاصه‌ای از سیستم‌های پژوهشی دسته‌بندی نتایج جستجو	55
جدول(2-10)	دسته‌بندی نتایج جستجو در مقابل دسته‌بندی‌های عمومی و سنتی متون	60
جدول(4-1)	نرخ به روز رسانی اخبار در سایت‌های پایه	82
جدول(4-2)	نمودار خبری مربوط به سه سازمان خبرگزاری پایه	84
جدول(4-3)	چکیده ساختار صفحه منبع کد سایت‌های پایه و سیستم نهایی	88
جدول(4-4)	تعداد اسناد آموزشی به تفکیک گروه‌ها	89
جدول(4-5)	نمودار گروه بندی سازمان‌های خبرگزاری و نمودار خبری نهایی	90
جدول(4-6)	یکسان‌سازی حروف مشابه	92
جدول(4-7)	تعداد اسناد آزمایشی به تفکیک گروه‌ها	98
جدول(4-8)	تشابه/ عدم تشابه گروه‌های خبرگزاری، انتسابی و انتخابی	102
جدول(4-9)	نمونه ریز متون تست و دسته‌های نسبت داده شده به آنها	103
جدول(4-10)	پارامترهای مربوط به معیارهای ارزیابی	105
جدول (4-11)	نتایج حاصل از ارزیابی سیستم نسبت به گروه انتخابی و خبرگزاری	107

فصل اول: مقدمه

1-1 پیشگفتار و ضرورت پروژه

تمایل به کسب اطلاع از اوضاع جامعه و پیگیری روند امور جاری به ویژه در محیط زندگی، از گذشته تاکنون یکی از اساسی‌ترین علاقه‌مندی‌های انسان‌ها را به خود معطوف داشته است. بر همین مبنا صنعت خبرگزاری به یکی از پرقدمت‌ترین صنایع در هر جامعه‌ای تبدیل شده است. امر خبرگزاری بسته به دوره زندگی و امکانات موجود در هر زمان به گونه‌ای متفاوت صورت پذیرفته است. در گذشته به دلیل گستره پراکندگی کم جمعیت، انتقال اخبار به صورت دهان به دهان از فردی به فرد دیگر انجام می‌شد، پس از مدتی، افراد مشخصی وظیفه این امر را به عهده گرفته و با اعلان همگانی در مراکز شهرها اهم اخبار را به گوش سایرین می‌رساندند. بعد از ایجاد صنعت چاپ و گسترش جغرافیایی شهرها ، اعلان‌های کاغذی جای نیروهای انسانی را گرفته و اخبار مهم به صورت مکتوب در اختیار شهروندان قرار داده می‌شد و عملا به جای یک فرد، یک گروه مشترک تحت عنوان یک سازمان خبرگزاری این وظیفه را به عهده گرفتند. وظیفه سازمان خبرگزاری تهیه و تدوین اخبار و قرار دادن آنها در دسترس عموم مردم است. این اخبار نه به معنای اعلامیه و یا اعلان بلکه تنها به صورت گزارشی از سیر وقایع اتفاقیه در زمینه‌های مختلف در اقصی نقاط جهان است که عموما به صورت روزنامه در اختیار مردم قرار داده می‌شود. در عصر جدید با ظهور و گسترش کاربردی تکنولوژی وب، امر اطلاع رسانی در این بستر و به صورت برخط، بلافاصله پس از اطلاع از وقوع حوادث صورت می-پذیرد. هرچند این روش مرزهای جغرافیایی و زمانی را شکسته و اطلاع‌رسانی را در کمترین زمان و با کمترین هزینه میسر می‌سازد، اما معضلات بسیاری را نیز به همراه آورده است.

با توجه به ماهیت پویای اطلاعات در صنعت خبرگزاری و به‌روزرسانی ثانیه‌ای اخبار در بستر وب در ساعات مختلف شبانه روز، نیاز به مراجعه مکرر به وب‌سایت‌های خبری برای مطلع شدن از جدیدترین اخبار وجود دارد. اولین راهکار استفاده از افرادی خبره به منزله دسته‌بندی اسناد و تیترهای خبری سازمان‌های خبرگزاری مختلف و بازپخش آنها در قالب گروه‌های مجزا است. از جمله مشکلات دسته بندی دستی مدارک توسط انسان، می‌توان به نیاز برای داشتن نیروی کار زیاد و متخصص در هر

زمینه، خطر تاثیرگذاری تصورات شخصی، عدم تشابه گروه‌بندی توسط دو انسان و در نتیجه زمانگیری و خطاپذیری بالا اشاره کرد.

مشکلاتی از قبیل ساعات کاری و نیروی انسانی محدود، تفاوت زمانی در نقاط مختلف جغرافیایی، تاثیر غیر مستقیم اخبار هر منطقه بر تحولات دیگر کشورها، سرعت بالای گسترش اخبار در اینترنت و عدم امکان کنترل آن پس از انتشار، افزایش چشمگیر منابع خبری با نمودارهای خبری متفاوت و حجم بالای اخبار منتشر شده در آنها در موضوعات و فیلدهای مختلف باعث سردرگمی و ابهام بیشتر در درک اخبار گردیده، تا آنجا که نیاز به یک سیستم برای متمایز ساختن داده‌ها از یکدیگر و دسته-بندی آن‌ها در گروه‌های متشابه در بستر وب به چشم می‌خورد.

نمایش و ارائه اخبار دسته‌بندی‌شده علاوه بر کاهش زمان اطلاع از اخبار، سهولت مطالعه و پیگیری عناوین مورد تمایل را نیز در برخواهد داشت.

2-1 تعریف پروژه

سرویس‌های خبری به صورت کلی به سه دسته خبرگزاری رسمی، پورتال اطلاع‌رسانی و سایت‌های خبری-تحلیلی تقسیم می‌شوند. هر چند تنها مورد اول به تولید و پخش خبر می‌پردازد، اما تعداد بالای این سازمان‌ها، روزانه بیش از صدها خبر مختلف را انتشار می‌دهد، تا جایی که می‌توان این چنین بیان نمود که تضاد عمیقی میان حجم اخبار و رشد سریع منابع قابل جستجو و توانایی در کنترل و زمانبندی ورود این اطلاعات به شبکه و توان، زمان و ساعات کاری نیروی انسانی وجود دارد و برای مقابله با آن، طراحی و پیاده‌سازی موتور جستجوگر و دسته‌بند اسناد خبری در وب ضروری است. در این پروژه با استفاده از مفهوم وب‌کاوی و به کمک یک خدمتگزار وب به استخراج و کشف خودکار اسناد خبری از وب سایت متعلق به خبرگزاری‌های رسمی می‌پردازیم. بدین منظور پس از دریافت اخبار، جهت ارائه اخبار در دسته‌های مورد نظر از تکنیک‌های دسته‌بندی استفاده نموده و هر کدام از اسناد را پس از انجام پیش پردازش‌های مورد نیاز اولیه و نهایی، داده کاوی مرتبط و استخراج ویژگی‌های مورد نظر، به یکی از شاخه‌های تعریف شده در نمودار خبری نهایی نسبت می‌دهیم.

هدف نهایی از انجام پروژه طراحی و پیاده‌سازی سیستمی برای دسته‌بندی خودکار اسناد خبری به دست آمده از سایت‌های اینترنتی جهت ارائه هرچه سریعتر و بهتر اطلاعات بوده و به عنوان یک سیستم میانی به تسهیل کار افراد می‌پردازد. هر چند در ادامه می‌توان از این روش به منزله شخصی سازی اسناد خبری برای کاربران صفحات خبرگزاری‌ها نیز استفاده نمود.

در ابتدا، کار خود را با تعداد محدود و نابرابر اسناد مربوط به هر دسته از خبرگزاری‌های پایه آغاز می-نماییم. در روند انجام کار باید به نکاتی همچون نحوه بدست آوردن اسناد خبری، نرمال و یکدست بودن متن‌های ورودی، روش استخراج و پردازش کلمات کلیدی و سنجش میزان تعلق کلمات به دسته‌ها و به‌روزرسانی پایگاه داده کلمات کلیدی پرداخت. عملا ما فرآیند استخراج متن را برای بکارگیری اطلاعات درون مدارک انجام می‌دهیم که برای این کار باید با جستجوی متن، الگوهای درون آنرا یافته و روابط میان کلمات با یکدیگر و ارتباط با کلمات در دسته‌های نهایی را بیابیم. از آنجا که ماشین به صورت خودکار این عمل را انجام می‌دهد، در نتیجه پروسه‌ای کامل برای یادگیری آن تعریف نموده‌ایم.

قسمت اول به طراحی موتور جستجوگر هماهنگ با سیاست‌های سازمان می‌پردازد. در این موتور به منزله کنترل صحت اخبار تنها از سایت‌های معرفی شده به سیستم استفاده می‌نماییم. یک خدمتگذار وب به صفحه کدینگ هر کدام از وب‌سایت‌ها رفته و با جستجو لینک‌ها و تشخیص لینک‌های مفید از نظر برنامه، اقدام به خواندن و ثبت محتوای آنها می‌کند. بدون شک یکی از چالش‌های پیش رو در این مرحله ایجاد تمایز میان اسناد خبری و غیرخبری موجود در صفحات است، که این عمل با تعریف مقادیر پیش‌فرض در عامل جستجوگر صورت می‌گیرد. به طور کلی هر خبر شامل سه مولفه اساسی تیتر، متن یا بدنه خبر و مرجع اولیه است، که در این پروژه تنها از بخشی از بدنه متن استفاده نموده و کل محتویات را مورد پردازش قرار نمی‌دهیم. در نهایت اسناد به دست آمده از این مرحله در پایگاه داده مربوطه ذخیره گشته و ویژگی‌های مورد نظر آن‌ها استخراج می‌گردد.

قسمت دوم کار اختصاص به مطالعه و تحلیل اسناد آرشیو خبرگزاری به منزله استخراج کلمات کلیدی مورد نظر در هر زیر شاخه و یافتن الگوی پر تکرار خبری دارد. در این بخش با علم بر شاخه‌های متناظر داده‌ها، کلمات کلیدی مربوط به هر بخش با وزن متناسب برای هر گروه بدست می‌آید. برای افزایش دادن جامعیت کار از نظر کاربران انسانی و دایرکتوری‌های معتبر نیز کمک گرفته و به صورت ترکیبی، کلمات حاصل از آنها را به سیستم اضافه می‌کنیم.

سپس با ادغام دو قسمت قبل فاز آموزش و تهیه داده اصلی مورد آزمایش اتمام یافته، برای تست سیستم به دریافت مجموعه خبری جدیدی رو می‌آوریم. برای سنجش میزان تشابه، کلمات کلیدی متن تست ورودی را با پایگاه‌های داده انحصاری گروه‌ها تطابق داده و گروه با بیشترین میزان شباهت را به عنوان گروه برنده انتخاب می‌کنیم. باید توجه داشت که در برخی موارد، سند خبری به بیش از یک گروه تعلق دارد. در انتها نیز بر اساس تمایل کاربر به مشاهده گروه‌های خبری خاص و یا تمام گروه‌ها، اسناد لیبل گذاری شده نمایش داده می‌شوند.

از مزایای این طرح می‌توان به استفاده بهینه از نیروی کار انسانی، به روز رسانی آنلاین اخبار بدون نیاز به حضور فیزیکی در محل، و در نهایت بهبود عملکرد سازمانی و کسب رضایتمندی کاربران اشاره کرد. تاکنون در چندین کشور تلاش برای دسته‌بندی خودکار اسناد موجود و ذخیره شده در پایگاه داده صورت گرفته است که از آن جمله می‌توان به سیستم آزمایشگاهی دسته‌بندی اخبار بر مبنای تیتر در دو کشور سوئد و کرواسی اشاره نمود، در کشور اسپانیا نیز پروژه ای مشابه تحت استاندارد کنسول بین المللی انتشار اخبار، IPTC[1] در حال تکمیل است. اقداماتی نیز جهت دسته‌بندی اسناد و آرشیوهای خبری در داخل کشور بر روی داده‌های به زبان فارسی صورت گرفته است که اکثر آنها بر روی داده‌های ثابت و غیرپویا آزمایش شده‌اند. باید به این نکته توجه داشت که در این پروژه تنها به موتورهای دسته بند اسناد که به معرفی الگوریتم‌ها و نحوه عملکرد پرداخته‌اند، اشاره نموده و تحلیل و مقایسه خروجی کار با سیستم‌های بسته تجاری مشابه در هیچ یک از بخش‌ها صورت نمی پذیرد.

[1] International Press Telecommunication Council

3-1 ساختار پروژه

در این فصل تنها به معرفی و نمایش رویکرد کلی، اهمیت، اهداف و سیر تکاملی پروژه، بدون توجه به ساختارهای فنی و جزئیات کار پرداخته شد. فصل دوم به معرفی مبانی نظری و پایه کار پرداخته و در مواردی به صورت تئوری مقایسه و تحلیل روش‌های مختلف آورده شده است. این فصل به ترتیب مباحث مربوط به موتورهای جستجوگر وب، سیستم‌های دسته‌بندی اسناد و نهایتا سیستم‌های تلفیقی به نام موتورهای جستجوگر دسته بند اسناد در وب را پوشش داده است.

در فصل سوم نگاهی انحصاری به پروژه‌ها و تحقیقات انجام شده در مباحث فصل دوم پرداخته و مروری بر اقدامات پیشین خواهیم داشت. فصل چهارم به شرح روش تحقیق و معرفی مراحل تکمیل کار پرداخته و به صورت تفصیلی جزئیات و معماری بکار گرفته شده در سیستم را نمایش می‌دهد. در انتها نیز ارزیابی سیستم صورت گرفته است.

در فصل پنجم پس از ارایه جمع‌بندی و تحلیل نتایج به دست آمده پروژه، راهکارهایی برای اقدامات آینده جهت ادامه پژوهش پیشنهاد شده است.

فصل دوم: مبانی نظری و پایه

1-2 مقدمه

با افزایش رو به رشد اطلاعات در بستر وب، یافتن سریع اطلاعات مورد نظر درخواستی بسیار دشوارتر از قبل گشته‌است. یکی از راهکارهای حل این معضل استفاده از تکنیک‌های دسته‌بندی اسناد در گروه‌های مشخص به منزله مدیریت داده‌ها، تسهیل در نمایش و ارائه نتایج به صورت منسجم است. در این بخش نگاهی کلی به دنیای وب و موجودیت‌های آن پرداخته و در ادامه مبحث دسته‌بندی اسناد خبری را مورد مطالعه قرار می‌دهیم. دسته بندی نتایج یکی از قابلیت‌هایی است که به موتورهای جستجوگر امروزی اضافه شده است، در حالی که تا سال‌ها تنها قابلیت موجود در آنها مرتب‌سازی اسناد بوده‌است. در ادامه با ترکیب دو مورد فوق تعریفی از موتورهای دسته‌بند اسناد را ارائه داده و کار را با محوریت داده‌های خبری و نیازمندی‌های خاص این حوزه پیش خواهیم گرفت.

2-2 دنیای وب

دنیای وب منبع عظیمی از اطلاعات است که روز به روز بر حجم آن افزوده می‌شود. در حال حاضر میلیاردها صفحه که اطلاعات فراوانی از موضوعات مختلف را دربردارند، بر روی سرورهای مختلف ذخیره شده‌اند. این در حالی است که تولید سایت‌های جدید و گسترش سایت‌های موجود نیز به صورت چشمگیری به حجم این اطلاعات می‌افزاید. در بررسی‌های گوناگون انجام شده در زمینه‌ی گسترش وب تخمین زده شده است که روزانه بیش از یک میلیون صفحه به وب اضافه می‌شود و بیش از 600 گیگابایت از صفحات در هر ماه تغییر می‌کنند[1] و[2] . نرخ رشد اطلاعات تا به آنجا رسیده که امروزه مشکل دسترسی به اطلاعات جدی‌تر از نبود اطلاعات است و چالش عمده کاربران، دستیابی مستقیم به اطلاعات مورد نظر است. به عبارت دیگر اگر کاربری دنبال موضوعی خاص باشد، کدام صفحه را باید بخواند؟ از میان این تعداد عظیم صفحات موجود، کدامین صفحه نیاز او را برآورده می-کند؟ در نتیجه باید ابزاری وجود داشته باشد که به کاربران در پیدا کردن اطلاعات کمک کند که ما آن‌ها را با نام موتور جستجوگر می‌شناسیم. هر چه بر محبوبیت وب افزوده می‌گردد، نیاز به بایگانی کردن اطلاعات آن نیز بیشتر می‌شود. از طرف دیگر داده‌ها به سرعت در حال تغییرند و رشد وب از

لحاظ تغییر اطلاعات و به روز شدن آنها در کسری از ثانیه رخ می‌دهد. به دلیل حجم بالا و امکان دسترسی عمومی سازماندهی بسیار بد و ناکاملی بر روی صفحات و اطلاعات مندرج بر روی آنها وجود دارد، در کل می‌توان چنین بیان داشت که بستر وب کاملاً غیر ساخت یافته است. نیازمندی‌ها کاربران وب نیز بسیار متفاوت بوده و هرکاربری ممکن است تنها به بخش کوچکی از وب علاقمند باشد، این امر مشکل یافتن اطلاعات موردنظر را چندین برابر می‌نماید. وجود این مشکلات کارشناسان را بر آن داشته است تا در حد ممکن به کاوش وب و کشف روابط و ساختارهای آن بپردازند. به صورت کلی وب کاوی به سه دسته مختلف زیر تقسیم می‌گردد که در هر کدام به دنبال روابط و خصوصیات مشخصی از وب هستیم.

1-2-2 وب کاوی

وب‌کاوی از مهمترین کاربردها در کاوش شبکهٔ گستردهٔ جهانی برای کشف و استخراج الگوهای مفید است. وب علاوه بر اینکه دارای مجموعه عظیمی از اطلاعات است، حاوی مجموعه‌ای پویا از پیوندها برای دسترسی به صفحات وب و استفاده از اطلاعات نیز می‌باشد که یک مجموعهٔ غنی برای داده‌کاوی ایجاد می‌کند. به طور کلی عملیات مورد نظر در وب‌کاوی شامل کاوش محتوایی وب، کاوش ساختاری وب و کاوش کاربردی وب است. باید توجه داشت که این شاخه‌های مختلف، کاملاً از هم مجزا نبوده و از این عملیات برای حل مشکلات مطرح شده در کاوش شبکهٔ گستردهٔ جهانی استفاده می‌شود. جدول (2-1) مقایسه بین عملیات موتورهای جستجوگر در هر کدام از جنبه‌های وب‌کاوی را نشان داده است.

کاوش محتوایی وب

کاوش محتوایی وب کشف اطلاعات مفید از اسناد داده و محتوای وب می‌باشد. آنچه با عنوان محتوای وب شناخته می‌شود، می‌تواند شامل انواع مختلفی از داده باشد. محتوای وب شامل داده‌های غیرساخت‌یافته مثل متون آزاد و داده‌های نیمه‌ساخت‌یافته مثل اسناد HTML و یک نوع داده ساخت‌یافته‌تر مثل داده موجود در جداول یا پایگاه‌داده‌های تولید کننده صفحات HTML است. در

هر حال بیشتر محتوای وب داده غیرساخت‌یافته است. تحقیقات دربارهٔ بکار بردن تکنیکهای داده کاوی از متون غیر ساخت یافته با عنوان متن‌کاوی [3] نامیده می‌شود. بنابراین می‌توانیم متن‌کاوی را به عنوان نمونه ای از کاوش محتوای متن بدانیم. ما می‌توانیم تحقیقات انجام شده در زمینه کاوش محتوایی وب را از دو نقطه نظر متفاوت بررسی کنیم: دیدگاه‌های بازیابی اطلاعات و پایگاه‌های داده. هدف کاوش محتوایی وب از دیدگاه بازیابی اطلاعات اساساً کمک به بهبود جستجوی اطلاعات یا فیلتر کردن اطلاعات برای کاربران و معمولاً بر اساس پروفایل‌های استنباط شده یا تقاضا شده کاربران است. در حالیکه هدف کاوش محتوایی وب از نقطه نظر پایگاه‌داده اساساً تلاش برای مدل کردن داده روی وب و استفاده از آن به نحوی است که بتوان به جای جستجوهای ساده که تنها براساس کلمات کلیدی هستند پرس‌وجوهای پیشرفته‌تری داشت[4].

کاوش ساختاری وب

کاوش ساختاری وب به دنبال کشف مدلی است که در زیرساختار پیوندهای وب وجود دارد. کاوش ساختاری وب می‌تواند برای کشف اعتبار صفحات وب استفاده شود. تعیین اعتبار صفحات وب در پیوندهایی که به آن صفحه وجود دارد، پنهان است. وب تنها شامل یک سری صفحات جدا از هم نیست بلکه مهمترین ویژگی آن پیوندهایی است که صفحات را به هم متصل می‌کنند. این پیوندها دارای مقدار زیادی اطلاعات نهفته هستند که در ایجاد این تصور که اعتبارسنجی صفحات به صورت خودکار انجام می‌شود، کمک زیادی می‌کنند. در واقع وقتی نویسندهٔ یک صفحهٔ وب پیوندی به صفحهٔ دیگری در صفحه‌اش می گذارد این پیوند به منزلهٔ مهر تأییدی است که این نویسنده روی آن صفحه زده است. هر چه یک صفحهٔ وب دارای پیوندهای بیشتری از جاهای مختلف باشد، این نشان‌دهندهٔ اهمیت بیشتر آن صفحه است و طبعاً اعتبار بیشتر آن صفحه را نشان می‌دهد. بنابراین پیوندهای وب اطلاعات غنی و کاملی در رابطه با مرتبط بودن، کیفیت و ساختار محتوایی وب ایجاد می‌کند و یک منبع غنی برای وب کاوی محسوب می‌شود. اما بر خلاف اظهار نظرهایی که ارائه می‌شود، پیوندهای وب ویژگی‌های خاصی دارند که استفاده از آنها را برای تعیین اعتبار صفحات با مشکل مواجه می‌کند.

اول اینکه هر پیوند نشان‌دهندهٔ تأییدی که ما به دنبال آن هستیم نیست. در واقع بعضی از این پیوندها به هدف دیگری در نظر گرفته شده‌اند. دوم اینکه ممکن است به دلایل تجاری یا رقابتی یک صفحه دارای پیوندی از طرف نویسندگان رقیب نباشد که این مسئله سبب پایین آمدن اعتبار آن صفحه می‌شود[4].

کاوش کاربردی وب

در کاوش کاربردی وب رکوردهای گزارشهای وب برای کشف الگوهای دسترسی کاربر به صفحات وب بررسی می‌شوند. در حالیکه کاوش محتوایی وب و کاوش ساختاری وب از داده اولیه و واقعی وب استفاده می‌کنند، کاوش کاربردی وب داده ثانویه مشتق شده از فعل و انفعالات کاربران با وب را می-کاود. تحلیل و پیدا کردن نظم موجود در رکوردهای گزارشهای وب می‌تواند مشتریهای احتمالی موجود در تجارت الکترونیکی را مشخص کند، کیفیت سرویس‌های اطلاعاتی اینترنت را برای کاربران افزایش دهد و کارایی سیستم سرویس دهندهٔ وب را بهتر کند[4].

فعالیت های موتورهای چستجوگر	کاوش محتوایی وب	کاوش ساختاری وب	کاوش کاربردی وب
دسترسی به اطلاعات مورد نظر کاربر از طریق موتورهای جستجو با وجود بزرگی و پویایی وب	روشهای متن کاوی ، بخش بندی اسناد وب، روشهای بازیابی اطلاعات		
افزایش کیفیت سرویسهای اطلاعاتی اینترنت		کشف اعتبار صفحات از طریق ساختار پیوندهای هر صفحه	کاوش گزارشهای دسترسی به سرویس دهنده های وب و متا اطلاعات
کاربر فقط بخشی از وب را که به مربوط اوست ببیند.	بخش بندی اسناد وب		کاوش گزارشهای دسترسی به سرویس دهنده های وب و متا اطلاعات
مدل کردن کاربر بر اساس علایق و اهداف او			کاوش گزارشهای دسترسی به سرویس دهنده های وب و متا اطلاعات
اعتبار دهی صفحات وب		کشف اعتبار صفحات از طریق ساختار پیوندهای هر صفحه	
سازماندهی مناسب وب	طبقه بندی صفحات وب	بخش بندی وب از طریق بررسی ساختار پیوندها بین صفحات مختلف	کاوش گزارشهای دسترسی به سرویس دهنده های وب و متا داده

جدول(2-1): نحوهٔ تشخیص عملیات مناسب برای حل مشکلات موتورهای جستجو [4]

ابزارها و روش‌های مناسب برای انجام هر یک از عملیات وب کاوی نیز در جدول (2-2) نشان داده شده است.

روشها و ابزار وب کاوی		تکنیکها	کاربرد
روشهای وب کاوی	روشهای متن کاوی	قوانین اپیزودی، روشهای پردازش زبانهای طبیعی، ابزارهای شبکه های عصبی، ابزارهای خوشه بندی	کاوش محتوای متون غیر ساختیافته وب برای بهبود کارایی موتورهای جستجو
	روشهای بازیابی اطلاعات	مدل مارکوف نهان، نقشه های خود سازمان ده، انتروژی بیشینه، روشهای استقرایی	کاوش محتوای اسناد وب برای بهبود کارایی موتورهای جستجو، پیدا کردن الگوهایی در متون اسناد وب
	روشهای بخش بندی متن	نزدیکترین Kهمسایه ها ، بردارهای پشتیبانی ماشین، روشهای استقرایی، سیستمهای خبره، شبکه های عصبی، یادگیری آماری	کاوش محتوای اسناد وب برای بهبود کارایی موتورهای جستجو
	کاوش قوانین انجمنی	قوانین انجمنی	مدل کردن کاربر، مشخص کردن شمای وب سایت
	روشهای طبقه بندی	درختهای تصمیم گیری، قانون بیز، الگوریتمهای ژنتیک، شبکه های عصبی	طبقه بندی اسناد وب
	روشهای خوشه بندی	روشهای جزء بندی، روشهای سلسله مراتبی، روشهای برمبنای تراکم، روشهای شبکه ای	خوشه بندی اسناد وب و تعیین بخشهای اولیه به منظور بخش بندی
ابزار	روشهای یادگیری آماری	رگرسیون، کوچکترین مربع خطی مناسب	بخش بندی اسناد وب، مدل کردن کاربر
	شبکه های عصبی	شبکه انتشار رو به عقب	بخش بندی اسناد وب

جدول (2-2): ابزارها و روش‌های وب کاوی [4]

2-2-2 روش های دسترسی به اطلاعات وب

پس از تعریف محیط وب و انواع کاوش‌های ممکن در آن، نوبت به معرفی ابزارها و راه‌های دسترسی به اطلاعات در این بستر است. بدون شک آسان‌ترین روش در این زمینه مراجعه به صفحه حاوی اطلاعات مورد نظر با علم بر آدرس آن است. اما به دلیل حجم بالا اطلاعات و تغییرات متناوب صفحات، امکان اطلاع از تمام آنها وجود ندارد، در نتیجه باید به خدمات جستجو در وب مراجعه نمود. خدمات جستجویی که در وب ارایه می‌شود را می‌توان در سه گروه اصلی فهرست (دایرکتوری)ها و موتورهای جستجوگر و دروازه‌های وب دسته‌بندی نمود.

تفاوت اصلی این گروه‌ها در نحوه تشکیل پایگاه داده و جمع‌آوری اطلاعات آنهاست. در فهرست‌ها، اینکار به عهده انسان است اما در موتورهای جستجوگر جمع‌آوری اطلاعات پایگاه داده را نرم افزارها انجام می‌دهند.

دایرکتوری[2]

ساختار دایرکتوری‌ها در واقع نوعی رده‌بندی عمومی را در اختیار می‌گذارند که از مفاهیم کلی و انتزاعی در سطوح بالای رده‌بندی شروع می‌شود و به تدریج با مفاهیم خاص گسترش پیدا می‌کند. برای هر یک از گره‌های این ساختار، تعدادی صفحه‌ی وب مرتبط با آن مفهوم، به آن گره نسبت داده شده است. همچنین به ازای هر گره، یک یا چند جمله وجود دارد که آن را توصیف می‌کند. از جمله معروف ترین این ساختارها می‌توان دایرکتوری‌های وب Dmoz و Yahoo را نام برد. این ساختارها توسط انسان ایجاد شده‌اند و بطور مکرر مثلا هر هفته یکبار به روز می‌شوند. فهرست هرگز از وجود سایت شما اطلاع نمی‌یابد مگر زمانی که شخصی آن را معرفی نماید. بعد از معرفی، ویراستار فهرست به سایت شما مراجعه نموده، در صورت رعایت قوانین فهرست و انتخاب گروه مناسب، سایت شما را به پایگاه داده فهرست اضافه می‌نماید. در واقع شما باید سایت خود را با عنوان و توضیحی مناسب به فهرست‌ها معرفی نمایید و بهترین گروه ممکن را برای سایت خود در نظر بگیرید. این کار بسیار مهم است زیرا عموما فهرست‌ها همین عنوان و توضیح را به همراه آدرس صفحه اول سایت‌تان در پایگاه داده خود قرار می‌دهند.

یکی از مشکلات این ساختارها این است که معمولا بصورت متوازن نیستند. به این معنی که گره‌های خاصی بسیار گسترده‌شده اند و گره‌های مربوط به برخی مفاهیم کمتر گسترش یافته‌اند. در نتیجه در بسیاری از موارد این ساختارها نمی‌توانند کمک چندانی داشته باشند.

مشکل عمده‌ی این ساختارها جهت استفاده از آن‌ها به عنوان یک منبع معنایی لغوی آن است که هر گره اطلاعات بسیار کمی در مورد مفهوم نسبت داده شده به آن در اختیار می‌گذارد و به منظور

[2] Directory

دستیابی به اطلاعات بیشتر باید عمل خزش وب را روی ساختار انجام داد که زمانبر است و به پردازش صفحات وب نیاز دارد. حتی در صورت انجام این کار، به دلیل نویز فراوان موجود در صفحات وب، دریافت اطلاعات در مورد آن مفهوم از وب‌سایت‌های نسبت داده شده به آن مشکل است. به علاوه مشکل پوشش نیز در این ساختار وجود دارد، به این معنی که لیست سایت های هر گره ممکن است تمامی جنبه های آن مفهوم را پوشش ندهند[5] .

دروازه وب[3]

ترجمه کلمه پرتال به فارسی در فرهنگ‌های لغت «دریچه»، «درگاه» و «مدخل» ذکر شده اما در تکنولوژی اطلاعات ، پورتال صفحه وب واسطی است که امکان دسترسی آسان را به هر چیزی که کاربر، برای انجام وظیفه یا خواسته‌اش نیاز دارد – بدون توجه به اینکه محل فیزیکی آن کجاست-، فراهم می‌کند. دروازه‌ای که به تبعیت از ذات اصلی خود یعنی همان دسترسی آسان، دسترسی ما را به وب و در نهایت اینترنت راحت‌تر از پیش می‌کند و محلی است برای به اشتراک گذاری خدمات و محتوا توسط چندین وب سایت قدرتمند . در واقع دروازه وب به پایگاه وبی گفته می‌شود که اطلاعات را با توجه به موضوع آنها دسته‌بندی می‌کند تا به این صورت به یافتن آنچه به دنبال آن هستیم، کمک کند [6].

وب پورتال یک محصول خاص نیست و همانند وب سایت یک ماهیت استاندارد و تشکیل شده از دو زیرساخت خدمات و محتویات است که این ماهیت، خروجی چندین استاندارد، دیتا و موتور خاص، زیر نظر یک یا چند مجموعه نظارتی مشخص و همگام است.

ویژگی‌های اصلی یک پورتال عبارت است از تجمع اطلاعات، هدف دار بودن اطلاعات، در دسترس بودن اطلاعات و ایجاد دریچه ورود منحصر به فرد. هدف وب پورتال‌ها طبقه‌بندی کردن اطلاعات و نیز تعریف دسترسی آسان به آنهاست و برای جلوگیری از پراکندگی در چگونگی یافتن، دسترسی و

[3] Portal

نگهداری اطلاعات توسط کاربر و به صورت خلاصه جلوگیری از سردرگمی آنها، به صورت همزمان سرویس‌های اصلی و جانبی را در اختیار کاربران قرار می‌دهند

پورتال همواره ما را به سایت‌ها یا پورتال‌های دیگر راهنمایی می‌کند و به خودی خود تنها یک راهنما است. برای همین است که در بعضی از موارد به پورتال‌ها، صفحات زرد اینترنتی[4] نیز می‌گویند . آنها عموما" حاوی مطالبی هستند که جنبه اطلاعات عمومی دارد. داده‌هایی که از منابع مختلف بر روی یک پورتال جمع‌آوری می‌شوند، معمولا دارای پراکندگی فراوانی هستند. به همین علت، در بسیاری از پورتال‌ها، ابزارهایی مانند دایرکتوری قرار داده می‌شود تا این اطلاعات را طبقه‌بندی نماید. از سوی دیگر داده‌های قرار داده شده بر روی یک وب سایت، اولا" از منابع محدودتری تامین می‌شوند و ثانیا" دارای پراکندگی زیادی نبوده و حول یک محور و موضوع مشخص می‌باشد. پورتال یک سیستم کاربر محور می‌باشد. تامین اطلاعات پورتال را می‌توان یک مرکز ارائه خدمات و اطلاعات اینترنتی دانست که بر چهار پایه اصلی انطباق‌پذیری[5]، اختصاصی کردن[6]، توصیه و معرفی کردن محتوا[7] و و خلاصه کردن مطالب[8] استوار است.

موتورهای جستجوگر[9]

موتور جستجو ابزاری است که امکان جستجوی اطلاعات در وب را فراهم می‌کند[7]. موتور جستجوگر می‌تواند از وجود سایت شما اطلاع یابد، اگر راه ورود آن فراهم شده باشد. در واقع نرم افزار موتور جستجوگر هر لحظه در حال وبگردی و به روز رسانی اطلاع قدیمی و همینطور افزودن اطلاعات جدید به پایگاه داده موتور جستجوگر است.

---

٤ Internet yellow pages
٥ Customization
٦ Personalization
٧ Recommendation
٨ Excerpt content
٩ Search engines

اولین نسل موتور جستجوگر در سال 1994 با WebCrawler and Lycos, کار خود را آغاز کرده و تنها به نمایش یک صفحه اطلاعات مرتبط از لحاظ محتوی و فرمت می‌پرداختند و بیشتر درخواست‌های اطلاعاتی را پوشش می‌دادند.

نسل دوم موتور جستجوگر در سال 1998 با شرکت گوگل شروع به کار کرده و عمل اولویت‌بندی نتایج را به سیستم اضافه نموده است. در این سیستم‌ها از اطلاعات خاص خارج از محتویات صفحه مانند آنالیز لینک‌ها، جریان داده مبتنی بر کلیک و ... نیز استفاده شده است. تمامی درخواست‌های اطلاعاتی و دنباله روی[10] در این نسل پوشش داده شده است.

نسل سوم در اوایل سال 2000 کار خود را با تلاش برای ادغام منابع چندگانه از شواهد و دربرگیری تمام درخواست‌های ممکن آغاز نموده است. انتظار می‌رود که در نسل آینده شاهد موتورهای جستجوگر دسته‌بند و معناگرا بوده و بتوانیم با استفاده عمومی از آنها، نحوه و قدرت دسترسی کاربران به اطلاعات مورد نیازشان را افزایش دهیم.

بازیابی مدرن اطلاعات وب بیشترین تلاش خود را معطوف به استفاده از نتایج کلاسیک موجود نموده و بیشتر دست به توسعه ابتکاری و ادغام روش‌های سنتی نموده است. همه موتورهای جستجو در زمان پاسخ‌گویی به جستجوهای کاربران، تنها در پایگاه داده‌ای که در اختیار دارند به جستجو می-پردازند و نه در وب! موتور جستجوگر به کمک بخش‌های متفاوت خود، اطلاعات مورد نیاز را قبلا" جمع آوری، تجزیه و تحلیل می‌کند، آنرا در پایگاه داده‌اش ذخیره می‌نماید و به هنگام جستجوی کاربر تنها در همین پایگاه داده می‌گردد.

اجزاء موتورهای جستجوگر

بخش‌های مجزای یک موتور جستجوگر شامل عنکبوت[11]، خزنده[12]، بایگانی کننده[13]، پایگاه داده[14] و سیستم رتبه‌بندی[15] می‌گردد، که در ادامه به شرح مختصر آنها می‌پردازیم.

---

۱۰ Navigation
۱۱ Spider
۱۲ Crawler
۱۳ Indexer

عنکبوت

اسپایدر یا روبات، به صفحات مختلف سر می‌زند، محتوای آنها را می‌خواند، لینکها را دنبال می‌کند، اطلاعات مورد نیاز را جمع‌آوری می‌کند و آنرا در اختیار سایر بخش‌های موتور جستجوگر قرار می‌دهد. اسپایدر کدهای HTML صفحات را می‌بیند.

اسپایدر، به هنگام مشاهده صفحات، بر روی سرورها رد پا برجای می‌گذارد. شما اگر اجازه دسترسی به آمار دید و بازدیدهای صورت گرفته از یک سایت و اتفاقات انجام شده در آن را داشته باشید، می‌توانید مشخص کنید که اسپایدر کدام یک از موتورهای جستجوگر صفحات سایت را مورد بازدید قرار داده است.

کار یک اسپایدر، بسیار شبیه کار کاربران وب است. همانطور که کاربران، صفحات مختلف را بازدید می‌کنند، اسپایدر هم درست این کار را انجام می‌دهد با این تفاوت که اسپایدر کدهای HTML صفحات را می‌بیند اما کاربران نتیجه حاصل از کنار هم قرار گرفتن این کدها را. اسپایدرها کاربردهای دیگری نیز دارند، به عنوان مثال عده‌ای از آنها به سایت‌های مختلف مراجعه می‌کنند و فقط به بررسی فعال بودن لینک‌های آنها می‌پردازند و یا به دنبال آدرس ایمیل می‌گردند.

خزنده

کراولر، نرم افزاری است که به عنوان یک فرمانده برای اسپایدر عمل می‌کند. آن مشخص می‌کند که اسپایدر کدام صفحات را مورد بازدید قرار دهد. در واقع کراولر تصمیم می‌گیرد که کدام یک از لینک-های صفحه‌ای که اسپایدر در حال حاضر در آن قرار دارد، دنبال شود. ممکن است همه آنها را دنبال کند، بعضی‌ها را دنبال کند و یا هیچ کدام را دنبال نکند . کراولر، ممکن است قبلا" برنامه‌ریزی شده باشد که آدرس‌های خاصی را طبق برنامه، در اختیار اسپایدر قرار دهد تا از آنها دیدن کند. دنبال کردن لینک‌های یک صفحه به این بستگی دارد که موتور جستجوگر چه حجمی از اطلاعات یک

[۱۴] Database
[۱۵] Ranker

سایت را می‌تواند (می‌خواهد) در پایگاه داده‌اش ذخیره کند. همچنین ممکن است اجازه دسترسی به بعضی از صفحات به موتورهای جستجوگر داده نشده باشد .

شما به عنوان دارنده سایت، همان طور که دوست دارید موتورهای جستجوگر اطلاعات سایت شما را با خود ببرند، می‌توانید آنها را از بعضی صفحات سایت‌تان دور کنید و اجازه دسترسی به محتوای آن صفحات را به آنها ندهید. موتور جستجو، به صورت معمول قبل از ورود به هر سایتی ابتدا قوانین دسترسی به محتوای سایت را (در صورت وجود) در فایلی خاص بررسی می‌کند و از حقوق دسترسی خود اطلاع می‌یابد. به عمل کراولر ، خزش نیز می‌گویند.

بایگانی کننده

تمام اطلاعات جمع‌آوری شده توسط اسپایدر در اختیار ایندکسر قرار می‌گیرد. در این بخش اطلاعات ارسالی مورد تجزیه و تحلیل قرار می‌گیرند و به بخش‌های متفاوتی تقسیم می‌شوند. تجزیه و تحلیل بدین معنی است که مشخص می‌شود اطلاعات از کدام صفحه ارسال شده است، چه حجمی دارد، کلمات موجود در آن کدامند، و ... . در حقیقت ایندکسر، صفحه را به پارامترهای آن خرد می‌کند و تمام این پارامترها را به یک مقیاس عددی تبدیل می‌کند تا سیستم رتبه‌بندی بتواند پارامترهای صفحات مختلف را با هم مقایسه کند. در زمان تجزیه و تحلیل اطلاعات، ایندکسر برای کاهش حجم داده ها از بعضی کلمات که بسیار رایج هستند صرفنظر می‌کند. کلماتی نظیر a ، an، the، WWW، is و ... . از این گونه کلمات هستند.

پایگاه داده

تمام داده‌های تجزیه و تحلیل شده در ایندکسر، به پایگاه داده ارسال می‌گردد. در این بخش داده‌ها گروه‌بندی، کدگذاری و ذخیره می‌شود. همچنین داده ها قبل از آنکه ذخیره شوند، طبق تکنیک‌های خاصی فشرده می‌شوند تا حجم کمی از پایگاه داده را اشغال کنند. یک موتور جستجوگر باید پایگاه داده عظیمی داشته باشد و به طور مداوم حجم محتوای آنرا گسترش دهد و البته اطلاعات قدیمی را هم به روز رسانی نماید. بزرگی و به روز بودن پایگاه داده یک موتور جستجوگر برای آن امتیاز

محسوب می‌گردد. یکی از تفاوت‌های اصلی موتورهای جستجوگر در حجم پایگاه داده آنها و همچنین روش ذخیره‌سازی داده‌ها در پایگاه داده است .

سیستم رتبه بندی

بعد از آنکه تمام مراحل قبل انجام شد، موتور جستجوگر آماده پاسخ‌گویی به سوالات کاربران است. کاربران چند کلمه را در جعبه جستجوی آن وارد می‌کنند و سپس با فشردن اینتر منتظر پـــاسخ می-مانند. برای پاسخگویی به درخواست کاربر، ابتدا تمام صفحات موجود در پایگاه داده که به موضوع جستجو شده، مرتبط هستند، مشخص می‌شوند. پس از آن سیستم رتبه‌بندی وارد عمل شده، آنها را از بیشترین ارتباط تا کمترین ارتباط مرتب می‌کند و به عنوان نتایج جستجو به کاربر نمایش می‌دهد. حتی اگر موتور جستجوگر بهترین و کامل‌ترین پایگاه داده را داشته باشد اما نتواند پاسخ‌های مرتبطی را ارایه کند، یک موتور جستجوگر ضعیف خواهد بود. در حقیقت سیستم رتبه بندی قلب تپنده یک موتور جستجوگر است و تفاوت اصلی موتورهای جستجوگر در این بخش قرار دارد. سیستم رتبه‌بندی برای پاسخ‌گویی به سوالات کاربران، پارامترهای بسیاری را در نظر می‌گیرد تا بتواند بهترین پاسخ‌ها را در اختیار آنها قرار دهد. الگوریتم، مجموعه‌ای از دستورالعمل‌ها است که موتور جستجوگر با اعمال آنها بر پارامترهای صفحات موجود در پایگاه داده اش، تصمیم می‌گیرد که صفحات مرتبط را چگونه در نتایج جستجو مرتب کند. در حال حاضر قدرتمندترین سیستم رتبه‌بندی در اختیار گوگل می‌باشد.

نحوه عملکرد موتور جستجوگر

طرز کار موتور جستجوگر به طور خلاصه بدین گونه است که ابتدا یک آدرس وب را می‌یابد، آن را دنبال می‌کند و به صفحه‌ای جدید می‌رسد. محتوای آن صفحه را می‌خواند و پارامترهای آن را مشخص می‌کند. به عنوان مثال تعداد کلمات متن آن صفحه، حجم و تاریخ به روزرسانی آن، بعضی از پارامترهای آن صفحه است. سپس پارامترهای تعیین شده را به همراه آدرس آن صفحه به بایگانی موتور جستجوگر ارسال می‌کند و این اطلاعات در آنجا پس از فشرده‌سازی، ذخیره می‌گردد.

حال اگر کاربرکلماتی را جستجو کند، موتور جستجوگر در پایگاه داده‌ای که تشکیل داده است ابتدا تمام صفحات مرتبط با موضوع جستجو شده را یافته و سپس مرتبط‌ترین اسناد را به عنوان اولین نتیجه جستجو و بقیه صفحات را بر اساس میزان ارتباط بعد از آن در اختیار کاربر قرار می‌دهد. به عبارت دیگر اگر تعداد نتایج جستجو 1000 مورد باشد، سایت رده اول مرتبط‌ترین و سایت رده 1000 کم ارتباط‌ترین سایت به موضوع جستجو شده می‌باشد.

اگر عبارت یکسانی در تمام موتورهای جستجوگر، جستجو شود هیچ کدام از آنها نتایج یکسانی را ارائه نمی‌دهند و با نتایج کاملا متفاوتی روبرو می‌شویم. تفاوت در ارائه نتایج جستجو در موتورهای جستجوگر از تفاوت آنها در الگوریتم (سیستم رتبه‌بندی) و بایگانی داده‌هایشان ناشی می‌شود. حتی اگر همه آنها از بایگانی داده یکسانی نیز استفاده کنند، بازهم نتایج جستجویشان متفاوت خواهد بود. هر موتور جستجوگری برای رده‌بندی صفحات وب، از الگوریتم خاصی استفاده می‌کند که منحصر به خودش بوده و فوق‌العاده محرمانه می‌باشد. الگوریتم نیز مجموعه ای از دستورالعمل ها است که موتور جستجوگر به کمک آن تصمیم می‌گیرد که سایت ها را چگونه در خروجی‌اش مرتب کند.

ضریب نفوذ مناسب به معنای حضور در موتورهای جستجوگر مهم و عمده، بایگانی شدن هر چه بیشتر صفحات سایت در پایگاه داده آنها و قرار گرفتن در صفحه های اول تا پنجم نتایج جستجوی آنهاست.

انواع موتورهای جستجو

انواع مختلفی از موتورهای جستجوگر وجود دارند که در ادامه به معرفی و شرح مختصری از آنها خواهیم پرداخت.

موتورهای جستجو عمومی

همانطور که پیش از این اشاره شد، حجم اطلاعات و داده های اینترنتی آنقدر زیاد و رو به افزایش است که کاربران برای یافتن اطلاعات دلخواه خود، نیازمند ابزارهایی (موتورهای جستجوگر) هستند که به آنها در پیدا کردن اطلاعات کمک کنند؛ موتورهای جستجو با ارائه خدمات جستجو علاوه بر

تسهیل کار کاربران، رقابتی میان وب سایتها و وبلاگها ایجاد کرده‌اند. تا جایی که مدیران سایتها، وب سایتها و وبلاگهای خود را بر اساس عملکرد موتورهای جستجوگر بهینه‌سازی می‌کنند.

از معروف‌ترین نمونه‌های خارجی این نوع موتور جستجو می‌توان به گوگل و یاهو اشاره کرد و از نمونه‌های داخلی، که برای فارسی زبانان برنامه‌ریزی شده است، می‌توان جس جو را نام برد.

موتورهای جستجوگر چندگانه[16]

با وجود محبوبیت بسیار بالا موتورهای جستجوگر نظیر گوگل، یاهو و ... به دلیل تفاوت الگوریتم‌های مورد استفاده، محدوده پوششی این صفحات همپوشانی نداشته و هر کدام تنها قادر به پویش بخشی از فضای وب هستند، لذا برای ایجاد دیدی وسیع و همه جانبه نیاز به ابزاری داریم تا در حد امکان به تمام زوایای یک موضوع بپردازد. این جستجوگرها به هنگام دریافت درخواست کاربر، نتایج موتورهای جستجوی مختلف را برای آن موضوع بررسی می‌کنند و بهترین نتایج را در اختیار کاربر قرار می‌دهند. در این نوع عمل پیمایش و ایندکس‌گذاری صفحات صورت نگرفته و بیشتر به تجمیع نتایج جستجوگرهای دیگر پرداخته می‌شود. در نهایت با حذف اطلاعات کاملا یکسان (محدوده پوششی مشترک)، به رتبه‌بندی مجدد اطلاعات پرداخته و آنها را به کاربر نمایش می‌دهند.

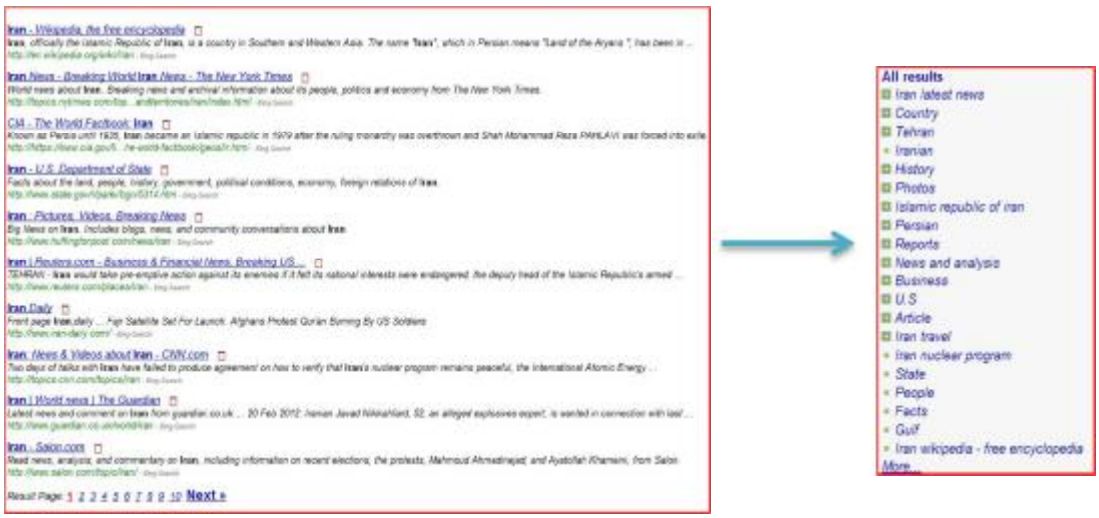
موتورهای جستجوگر خاص

این نوع موتورهای جستجو بر موضوعی خاص تمرکز دارند و تنها به جستجو به وب سایتها و وبلاگ-های مرتبط با آن موضوع می‌پردازند. یکی از نمونه‌های این نوع موتور جستجو Chemical Search می‌باشد. در این موارد حجم کلی فضای مورد جستجو را با آموزش دادن به جستجوگر کاهش داده و تنها مطالب مرتبط را به نمایش می‌گذاریم. از آنجا که در هر حوزه تعداد وب سایت‌های معتبر محدود هستند، در ابتدا به صورت پیش فرض از آنها استفاده می‌کنیم.

[16] Meta search engine

موتورهای دسته‌بندی نتایج وب

با وجود اینکه موتورهای جستجوگر چندگانه تا حد زیادی مشکل کاربران در دسترسی به منابع اطلاعاتی را کاهش داده‌اند، اما همچنان وجود تعداد نتیجه بازگردانده شده بالا موجب سردرگمی و ابهام در بدست‌آوردن اطلاعات است. از طرف دیگر با ادغام چندین موتور حجم بالایی از اطلاعات تکراری و مشابه در لیست نهایی مشاهده می‌گردد. عمل دسته بندی در واقع ساختار نمایش و ارائه اطلاعات را تغییر داده و هیچ گونه دخالتی در محتوای اسناد ندارد. تفاوت ارایه نتایج به صورت لیست ترتیبی و گروه‌های سازماندهی شده در شکل (2-1) مشاهده می‌گردد. به عنوان نمونه ای از این دسته می‌توان به موتور iboogie اشاره نمود.



شکل (2-1): تغییر ساختار ارائه اسناد وب از حالت لیست رتبه بندی شده به دسته های مجزا
{http://www.iboogie.com/searchtree.asp?name_query=iran&name_tab=0&name_do_search=1}

هدف از طراحی موتورهای دسته بند نتایج وب ارائه بهتر و منسجم‌تر داده‌ها به گونه ای است که سرعت کاربر در یافتن اطلاعات مورد نیاز و مورد علاقه‌اش را افزایش دهد. این امر علاوه بر صرفه‌جویی در هزینه و زمان، میزان رضایت کاربر را نیز بالاتر می‌برد. در مورد این موتورها به صورت مفصل در بخش سوم توضیح داده شده است.

3-2-2 کدهای وضعیت HTTP[17] و خطاهای سرور

هنگامی که در بستر پروتکل HTTP در وب در حال گشت و گذار و مرور صفحات مختلف هستیم، با وارد نمودن هر آدرسی که از طریق مرورگر خود از سرور سایت‌های مختلف درخواست می‌کنیم، کدهایی در پس زمینه بین واسط کاربری ما (مرورگر) و سرور رد و بدل می‌شوند که به آنها در اصطلاح کدهای وضعیت HTTP[18] می‌گویند، این کدها توسط کنسرسیوم‌ها و بنیادهای جهانی استاندارد سازی وب از جمله IETF[19] و W3C[20] تعریف و سازمان‌دهی شده‌اند و امروزه تقریبا سرور یا مرورگری وجود ندارد که از اصول آنها پیروی نکند. آشنایی با این کدها و مدیریت درست و به‌جا آنها باعث جلوگیری از ایجاد خطا در روند اجرا برنامه خواهد شد. این کدها به پنج دسته کلی تقسیم می‌شوند که در ادامه به آنها اشاره نموده و در فصل چهارم به تفصیل پیرامون کدهای چالش‌زا بحث خواهیم نمود.

کدهای سری 100، مربوط به اطلاعات

اولین سری از کدهایHTTP ، با عدد 100 شروع می‌شود که در مورد نقل و انتقال بسته‌های اطلاعات مثل ارسال و دریافت فایل، کاربرد دارند و حالت موقت پاسخ سرور را نشان می‌دهند، به فرض وقتی از متد POST در فرم‌های وب استفاده می‌کنیم، دریافت کد 100 به معنی این است که سرور درخواست ما را پذیرفته و فرایند پردازش اطلاعات ادامه دارد، االبته بدون ارسال کد 100 نیز این فرایند ادامه می‌یابد لذا ارسال آن از طرف سرور ضروری نیست و حتی در مرورگرهایی که از نسخه HTTP/1.0 استفاده می‌کنند، این کد قابل فهم و پردازش نیست.

[17] Hypertext Transfer Protocol
[18] HTTP response status codes
[19] Internet Engineering Task Force
[20] World Wide Web Consortium

کدهای سری 200، درخواست موفق

کدهای سری 200، به این معنی است که درخواست ارسالی واسط کاربری (که می‌تواند مرورگر یا ابزار دیگری باشد)، با موفقیت دریافت، موافقت، پردازش و پاسخ داده شده است، کدهای سری 200 معمولا به معنی بی نقص بودن درخواست و عملکرد صحیح سرور است.

کدهای سری 300، انتقال

کدهای سری 300 مربوط به مواردی هستند که پاسخ به درخواست واسط کاربری از سرور، باید با انجام اعمال دیگری (در سمت کاربر) کامل شود، این عملیات معمولا توسط واسط کاربری (مثلا مرورگر) و بدون دخالت کاربر (به صورت خودکار) انجام می‌شود، به فرض عمل انتقال خودکار از یک آدرس به آدرس دیگر، با ارسال کدهای سری 300 انجام می‌شود، نکته مهم در اینجا این مسئله است که انتقال‌ها نباید در یک درخواست، بیش از 5 بار تکرار شوند، در غیر اینصورت در اکثر مرورگرها، فرض بر حلقه (Loop) بی‌انتها شده و ارتباط قطع خواهد شد.

کدهای سری 400، خطای سمت کاربر

کدهای سری 400 مربوط به رویداد خطایی از جانب کاربر (سمت کاربر) در ارائه درخواست به سرور است، در پاسخ، سرور معمولا و به طور پیش فرض، به همراه کدهای HTTP عباراتی در توضیح خطای رخ داده ارسال می‌کند و دائمی یا موقتی بودن مشکل به وجود آمده را نیز تعیین خواهد کرد.

کدهای سری 500، خطای سمت سرور

کدهای سری 500 به معنی نقص داخلی سرور است، با این حال سرور در مجموع سالم بوده و احتمالا به طور موقت در حال انجام به روزرسانی یا تغییراتی است و در ساعات آینده مشکل رفع خواهد شد.

3-2 دسته‌بندی اسناد

فرآیند کشف زیر مجموعه‌هایی از اسناد ورودی در راستایی که اسناد داخل یک گروه بیشترین میزان شباهت را با یکدیگر داشته و در عین حال دارای کمترین تشابه با دیگر گروه‌ها باشند را دسته‌بندی می‌نامیم. میزان این تشابه بر مبنای معیارهای اندازه‌گیری عددی سنجیده می‌شود. در مورد اسناد متنی، تشابه باید میان عناصر تشکیل‌دهنده یا به عبارتی محتوای هر سند استوار باشد که عملا شامل کلمات و عبارات تشکیل‌دهنده آنها است. متن‌کاوی، به فرآیند کشف الگو در مجموعه متون غیر-ساخت‌یافته بر مبنای تکنولوژی‌های پردازش زبان طبیعی[21]، بازیابی اطلاعات[22]، استخراج اطلاعات[23] و داده‌کاوی[24] گفته می‌شود[8]، که در آن دو متن با تعداد بالای کلمات یکسان متشابه نامیده می‌شوند. با توجه به نحوه اختصاص اسناد به دسته‌ها و امکان ایجاد همپوشانی میان آنها، روش‌های دسته‌بندی را به دو گروه سخت[25] با قابلیت تعلق تنها به یک دسته و نرم[26] به منظور اختصاص به چندین گروه دسته‌بندی می‌کنیم. هر چند اکثر الگوریتم‌ها در دسته‌بندی اسناد بر مبنای روش اول عمل می‌کنند. از نقطه نظر نظارتی فرآیند گروه بندی را به دو مدل دسته‌بندی و خوشه‌بندی طبقه‌بندی می‌کنیم.

1-3-2 خوشه‌بندی[27]

خوشه‌بندی به عمل یادگیری بدون ناظر گفته می‌شود. در خوشه‌بندی هیچ نظارت یا اطلاعاتی به جز داده‌های آموزشی در اختیار یادگیرنده قرار ندارد و این یادگیرنده است که باید در داده‌ها به دنبال ساختار یا الگویی خاص بگردد. نتیجه خوشه‌بندی به نحوه نمایش آبجکت‌ها، معیار تشابه و الگوریتم آن بستگی دارد. در این روش به صورت خودکار ویژگی‌های متمایز کننده زیرگروه‌ها تعریف می‌گردد. انتظار می‌رود در نهایت به دسته‌هایی با عناصر درونی مشابه و متمایز با عناصر دیگر گروه‌ها دست یابیم [9,10,11].

---

[۲۱] Natural Language Processing
[۲۲] Information Retrieval
[۲۳] Information Extraction
[۲۴] Data Mining
[۲۵] Hard/ crisp
[۲۶] Soft/fuzzy
[۲۷] Clustering

در خوشه بندی نه تعداد گروه‌ها و نه خصوصیت یا ویژگی مورد نظر برای تمایز گروه‌ها از قبل مشخص نمی‌باشند. در نتیجه پس از انجام خوشه‌بندی لازم است یک فرد خبره، خوشه‌های ایجاد شده را تفسیر کند و در بعضی مواقع نیاز است که پس از بررسی خوشه‌ها، بعضی از پارامترهایی که در خوشه‌بندی در نظر گرفته شده‌اند ولی نامرتبط بوده و یا اهمیت چندانی ندارند، حذف گردیده و جریان خوشه‌بندی مجددا از اول صورت گیرد. پس از این که داده‌ها به چند گروه منطقی و توجیه-پذیر تقسیم شدند، از این تقسیم‌بندی می‌توان برای کسب اطلاعات در مورد داده‌ها یا تقسیم داده‌های جدید استفاده نمود. انواع الگوریتم‌های خوشه‌بندی را می‌توان به دو روش چند بخشی[28] و سلسله مراتبی[29] تقسیم نمود. با توجه به این که روش‌های خوشه‌بندی سلسله مراتبی، اطلاعات بیشتر و دقیق‌تری تولید می‌کنند، برای تحلیل داده‌های با جرئیات بیشتر، پیشنهاد می‌شوند، ولی از طرفی چون پیچیدگی محاسباتی بالایی دارند، برای مجموعه داده‌های بزرگ، روش‌های خوشه‌بندی مسطح پیشنهاد می‌شود. تمامی الگوریتم‌های خوشه‌بندی به صورت خلاصه در جدول (2-3) آورده شده‌اند. در ادامه به شرح هر کدام در حد نیاز پرداخته شده است.

[۲۸] Partitional
[۲۹] Hierarchical

Algorithm	Time complexity	Overlap	Clustering type	Describtion
Single linkage	$O(n^2)$	Hard	Hierarchial	n= number of documents
Group average	$O(n^2)$	Hard	Hierarchial	n= number of documents
Complete linkage	$O(n^2)$	Hard	Hierarchial	n= number of documents
Centroid/ Median HAC	$O(n^2)$	Hard	Hierarchial	n= number of documents
k-Means	O(nkt)	Hard	Partitional	n= number of documents k= number of initial clusters t= number of iterations
Single pass	O(nlogn)	Hard	Partitional	n= number of documents
Scatter/ gather	O(kn)/O(nm)	Hard	Partitional/ Hierarchial	n= number of documents k= number of initial clusters t= number of iterations
STC	O(n)	Soft	Hierarchial	n= length of all documents
SHOC[12]	O(n)	Soft	Hierarchial	n= length of all documents
Lingo[13,14]	$O(m^2n + n^3)$	Soft	Hierarchial	n= number of documents m= number of features

جدول(2-3): خلاصه ای از روش‌های مورد استفاده در خوشه‌بندی اسناد

روش چند بخشی

روش متداول در مدل مسطح kmeans [15] است که مرکز دسته ها را به عنوان centroid و نقطه مرکزی را به عنوان نماینده مناسبی از گروه در نظر گرفته و با انتخاب k مرکز دسته و مقایسه فاصله cosine یا هر معیار تشابه دیگری (در ادامه معیارهای تشابه توضیح داده شده است) میان اسناد و این مراکز، تعلق آنها را مشخص می‌کند. به روز رسانی مرکز پس از انتساب های جدید صورت گرفته و تکرار روند، تا زمانی که دیگر هیچ تغییری در گروه‌ها رخ ندهد ادامه می‌یابد. از معایب این روش می-توان به وابستگی بالا آن به انتخاب خوشه‌های اولیه و تعداد آن‌ها و عدم وجود روال مشخص برای محاسبه اولیه مراکز خوشه‌ها نام برد. هرچند در ادامه انواع مختلفی از روش‌ها با تغییرات جزئی در

الگوریتم اصلی توسط محققان پیشنهاد شد که در هر یک سعی در کم کردن وابستگی به مقداردهی اولیه است [16].

روش دیگر استفاده از سیستم پراکنده سازی- جمع آوری[30] است که از دو الگوریتم چند بخشی با پیچیدگی زمانی خطی استفاده می‌کند. شکستن هر دسته و جمع‌آوری مجدد زیر دسته‌ها عمل پالایش را در هر دسته انجام داده و از روش‌های بالابه پایین، یا پایین به بالا بسته به نوع داده به منظور انتخاب مراکز اولیه خوشه‌ها استفاده می‌گردد. هدف اصلی این روش انتخاب بهینه مراکز و سپس انتساب اسناد به آنها است.

آخرین الگوریتم مورد استفاده، روش مسیر ساده[31] نام دارد که در آن هر سند به مرکزی اختصاص می‌یابد که میزان تشابه میان آنها بیشتر از یک حد آستانه باشد. دسته ها با یک بار مرور اسناد ایجاد شده و هیچ تکراری صورت نمی‌گیرد.

مزیت اصلی این روش‌ها سهولت و پیچیدگی زمانی پایین آنها است. در حالی که عیب اصلی دسته-بندی نیاز به تعریف پارامترهای بسیاری مانند تعداد نهایی دسته‌ها، انتخاب اولیه مراکز دسته‌ها و ... است.

روش سلسله مراتبی

خروجی متدهای سلسله مراتبی، ساختاری درخت مانند دارد. این روش می‌تواند به دو صورت متراکم-ساز (هر سند یک دسته) و تفرقه‌ساز (تمام اسناد یک دسته منفرد) عمل کند. در روش تفرقه‌ساز ، ابتدا تمام داده‌ها به عنوان یک خوشه در نظر گرفته می‌شوند و پس از طی یک فرایند تکراری در هر مرحله، داده‌هایی که شباهت کمتری به هم دارند، به خوشه‌های مجزایی شکسته می‌شوند و این روال تا رسیدن به خوشه هایی که دارای یک عضو هستند، ادامه پیدا می‌کند. اما در روش متراکم ساز ابتدا هر داده به عنوان خوشه‌ای مجزا در نظر گرفته می‌شود و در طی فرآیندی تکراری در هر مرحله،

[30] Scatter/Gather
[31] Single Pass

خوشه هایی که شباهت بیشتری با یکدیگر دارند، ترکیب می‌شوند تا در نهایت یک خوشه و یا تعداد معینی خوشه حاصل شود. برخی از این الگوریتم‌ها عبارتند از [17]:

اتصال منفرد[32] : در این متد تشابه میان دو گروه برابر است با ماکزیمم تشابه میان تمام زوج اسناد، به گونه ای که هر عضو از دوتایی متعلق به یک گروه مجزا باشد. هر عضو از مجموعه باید حداقل به یکی از اعضا دسته شباهتی بیشتر نسبت به اسناد هر گروه دیگری داشته باشد. این روش دارای کمترین پیچیدگی زمانی است اما در ارائه نتایج مفید ضعیف عمل می کند.

اتصال کامل[33]: در این روش تشابه میان دو گروه برابر است با مینیمم تشابه میان تمام زوج اسناد، به گونه‌ای که هر عضو از دوتایی متعلق به یک گروه مجزا باشد. هر عضو از مجموعه حداقل به یکی از غیر مشابه‌ترین اعضا دسته شباهتی بیشتر نسبت به غیر مشابه‌ترین سند هر گروه دیگری دارد. در این روش خوشه‌هایی منسجم‌تر از روش فوق ایجاد می‌گردد.

میانگین اتصال گروهی[34]: در این متد تشابه میان دو گروه برابر است با میانگین تشابه میان تمام زوج اسناد، به گونه‌ای که هر عضو از دوتایی متعلق به یک گروه مجزا باشد و دارای بهترین نتیجه نسبت به دو مورد قبل می‌باشد.

روش‌های میانگین/ میانه[35]: هر گروه ایجاد شده توسط میانگین/ میانه اعضا نمایش داده می‌شود، در هر مرحله دسته‌هایی که تشابه بیشتری در میانگین/ میانه دارند، با هم ادغام می‌شوند. تفاوت دو معیار بررسی شده در آن است که مقدار میانه وابستگی به سایز دسته نداشته و به صورت مستقل از تعداد اعضا عمل می‌نماید.

روش درخت پسوندی[36]

این الگوریتم[18] با هدف دسته‌بندی اسناد متنی بازگردانده شده از موتورهای جستجوی وب بر اساس عبارات مشترک میان آنها طراحی شده است. ایده اصلی استفاده از عبارات به جای کلمات

---

[۳۲] Single link
[۳۳] Complete link
[۳٤] Group average link
[۳٥] Centroid/median methods
[۳٦] Suffix Tree Clustering

نشأت گرفته از میزان قدرت توصیفی بالا در آنها است. در این روش تمام اسناد وب به عنوان رشته متنی در نظر گرفته می‌شوند. دو مرحله کلی در این الگوریتم عبارتند از فاز تشخیص دسته پایه فاز ادغام دسته‌های پایه. در مرحله اول از یک درخت پیشوندی عمومی با استفاده از کلمات به عنوان المان‌های جملات استفاده شده است. گره‌های درخت نشان‌دهنده اسناد با عبارات یکسان بوده و مواردی که از حد آستانه مشخص بیشتر باشند، با هم ادغام می‌گردند. مهم‌ترین مزیت این روش استفاده از عبارات به جای کلمات برای ساخت برچسب‌های مفهومی‌تر است. نیاز به تعریف حد آستانه مانند تمام موارد دیگر مشکل عمده این روش می‌باشد، از طرف دیگر اگر سندی شامل عین عبارات تعیین شده نباشد، حتی در صورت تشابه با گروه هم امکان ادغام با آنها را ندارد.

روش آنلاین سلسله مراتبی مفهومی[37]

این الگوریتم که به روش SHOC شناخته شده است نیز برای دسته بندی نتایج وب و بر اساس روش آنالیز معنایی پنهان[38] برای زبان چینی طراحی شده است. تحلیل معنایی پنهان اولین بار درسال 1988[19] معرفی گردید. این روش سعی در کاهش بعد و بهبود معنای مدل برداری دارد و در آن، بردارهای عبارات به مفاهیم سطح بالاتر و در فضای با بعد کمتر نگاشت می‌شوند.

در تحلیل معنایی پنهان، همه‌ی مستندات وکلمات کلیدی، یک ماتریس دو بعدی مستند-کلمه را تشکیل می‌دهند. سپس روش تجزیه مقدار منفرد[39] (SVD) برای به دست آوردن ویژگی‌های معنایی، ماتریس مستند-کلمه را تجزیه می‌کند. روش SVD به منظور حذف اطلاعات نامربوط ماتریس مستند-کلمه با ابعاد زیاد و تبدیل آن به یک ماتریس مستند-کلمه با ابعاد کم استفاده می‌گردد. این روش بر خلاف STC از ساختار سلسله مراتبی و آرایه ای به جای درخت استفاده می‌کند. دو مفهوم کلی در این روش معرفی شده است که عبارتند از عبارت کامل و تعریف خوشه مستمر. در مورد اول به صورت کلی به دنبال ماکزیمم نمودن طول دنباله کلمات متوالی کشف شده در متن و اجتناب از

[37] Semantic hierarchical online clustering
[38] Latent Semantic Indexing
[39] Singular Value Decomposition

انتخاب عبارات کوتاه که به سختی قابل درک هستند، می‌باشیم. برای این کار از ساختار آرایه ای با پیچیدگی زمانی $O(n)$ که در آن n طول کل اسناد قابل پردازش است استفاده می‌کنیم. مورد دوم به اسناد اجازه می‌دهد تا با درجه عضویت متفاوت به چندین گروه تعلق داشته باشند. این مورد مانند روش فازی عمل نموده و مشکل هم‌پوشانی داده‌ها در گروه‌ها را برطرف می‌سازد. الگوریتم در سه فاز تشخیص عبارات کامل، کشف خوشه‌های پایه و همچنین ادغام خوشه‌های مشابه به کمک ساختار سلسله مراتبی عمل می‌کند. در فاز دوم از روش تجزیه مقدار منفرد به کمک برداهای متعامد و تعیین مقادیر ویژه استفاده می‌شود. وجود ابهام در تعیین مقادیر آستانه و مشاهده چندین نتیجه که تا حدی به صورت رندوم به تولید عبارات پرداخته بودند، از جمله معایب این روش محسوب می‌گردد.

روش لینگو[40]

معنای لغوی Lingo عبارت است از طیف وسیعی از کلمات یا سبکی از زبان که در موقعیت خاص یا توسط گروهی مشخص از مردم مورد استفاده قرار می‌گیرد. هر دفعه که کاربر پرس‌وجویی را در سایت مطرح می‌کند، یک زبان جدید با ویژگی‌های خاص خود مانند لغات، اصطلاحات و عبارات مختص آن ایجاد می‌شود. کاربران به احتمال زیاد نسبت به توصیفات بیش از حد طولانی و مبهم بی‌اعتنا خواهند بود. هرچند اطلاع داشته باشند که محتوای آنها ممکن است ارزشمند باشد. در نتیجه باید سعی در ایجاد برچسب هایی معنادار، مجزا و دقیق برای دسته ها نماییم. تعیین برچسب در روشهای متداول و مرسوم که بعد از تعیین گروه‌ها و بر اساس محتوای اسناد صورت می‌گیرد نسبتا دشوار است. روش-های عددی تا حد بالایی "می‌دانند" که اسناد مشابه یکدیگر هستند، اما قادر به شرح این تشابه نیستند. در روش لینگو ابتدا توصیفات مد نظر قرار گرفته و انتخاب دقیق برچسب‌های نامزد اهمیت بسیاری دارد. از طرف دیگر باید این اطمینان حاصل گردد که برچسب‌ها به اندازه کافی نسبت به یکدیگر متفاوت بوده و تمام فضای نتایج جستجو را پوشش داده‌اند. برای یافتن این کاندیدها از مدل

[40] Lingo

فضای بردار[41](VSM) و تجزیه مقدار منفرد استفاده شده است. یعنی در مرحله اول به ایجاد لیبل پرداخته و در ادامه تنها قطعات به دست آمده را به هر برچسب نسبت می‌دهیم. این الگوریتم در دو فاز کلی استقرا برچسب خوشه به کمک تکنیک کاهش ابعاد و فاز تخصیص محتوای خوشه توسط مدل فضای بردار عمل دسته‌بندی را انجام می‌دهد. متد VSMیکی از روشهای موجود در بازیابی اطلاعات است که از عملیات جبری خطی برای مقایسه داده‌های متنی استفاده می‌کند. این روش به هر یک از اسناد موجود در مجموعه، برداری چند بعدی اختصاص می‌دهد و هر قسمت از بردار به یکی از لغات موجود در سند اشاره دارد. در نتیجه هر سند به کمک ماتریس (لغت- سند) منحصر به فرد نمایش داده می‌شود. مقدار هر یک از اجزا در ماتریس بستگی به نوع و میزان وابستگی لغت و سند متناظر دارد.

برخلاف VSM، LSI مجموعه ورودی را به کمک مفاهیم موجود در اسناد به جای معنی واقعی کلمات موجود نمایش می‌دهد. برای این کار با استفاده از تعداد معدودی از عوامل متعامد ماتریس (لغت- سند) تخمین زده می‌شود. این فاکتورها به نمایش مجموعه‌ای از مفاهیم انتزاعی می‌پردازند، که هر کدام بخشی از ایده مشترک میان زیر مجموعه‌ای از اسناد کلی را بیان می‌دارد. از دیدگاه lingo این مفاهیم، بهترین کاندیدهای برچسب دسته هستند، هر چند فهم نمایش ماتریسی آنها برای انسان سخت است. برای بدست آوردن برچسب‌های خلاصه و حاوی اطلاعات مفید، باید بار معنایی و مفهومی لغات مورد پردازش در LSI را تا حد امکان افزایش دهیم. در بین تمام رخدادهای لغوی در سند، آنهایی که به تعداد دفعات بیشتر از حد آستانه در سند تکرار شده‌اند، ارزش معنایی بیشتری برای کاربر داشته و در عین حال پیدا کردن آنها آسان است. این مجموعه اغلب شامل موجودیت‌های اسمی است. الگوریتم lingo از این عبارات تکرار شونده به عنوان برچسب دسته‌های خود استفاده می‌کند. سپس توسط روش SVM اسناد را به دسته‌های مربوطه الحاق می‌کنیم. پیچیدگی زمانی این الگوریتم برابر با $O(m^2n+n^3)$ است[20] .

[41] Vector space model

2-3-2 دسته‌بندی

دسته‌بندی یا خوشه‌بندی بانظارت [9]، یک تکنیک داده‌کاوی است که از طریق ارزیابی ویژگی‌های مجموعه‌ای از داده‌ها، آنها را به تعدادی دسته‌های از پیش تعیین شده، اختصاص می‌دهد .این عمل گامی بلند در بازیابی، مدیریت و کاوش داده‌های موجود در سیستم‌های اطلاعاتی برداشته است.

خوشه بندی اسناد متنی به خوشه‌های مختلف، یک گام مهم در ایندکس کردن، بازیابی، مدیریت و کاوش داده‌های متنی فراوان موجود روی وب یا سیستم اطلاعات یک شرکت است. در این روش، عامل یادگیر با استفاده از یک سری از داده های آموزشی یاد می‌گیرد که در هر موقعیت چه کاری انجام دهد.

مسائل دسته‌بندی به شناسایی خصوصیاتی منجر می‌گردد که همزمان به درک مفاهیم موجود و پیش‌بینی نمونه‌های جدید کمک می‌کند. این روش اسناد را بر طبق یک ساختار شناخته شده مرتب می‌سازد. از مهم‌ترین روش‌های دسته‌بندی می‌توان به بیزین ساده، k- نزدیک‌ترین همسایه و ماشین بردار پشتیبان اشاره نمود.

روش بیزین

این روش متداول‌ترین روش در دسته‌بندی متون است که در آن متن به صورت مجموعه ای از کلمات مستقل از یکدیگر و مستقل از محل قرار گرفتن در متن دیده می‌شود. تعریف تابع احتمال هر متن از حاصلضرب احتمال کلمات آن و احتمال رخداد متنی با آن طول بدست می‌آید(2-1). احتمال هر دسته نیز از تعداد متن‌های متعلق به آن دسته نسبت به کل متن‌ها بدست می‌آید(2-2).

$$p(d_i) = p(|d_i|) * \prod_{k=1}^{|d_i|} p(w_k^i) \qquad (1-2)$$

$$p(c_j) = \frac{\#(d_i \in c_j)}{\#(d_i)} \qquad (2-2)$$

که در اینجا d_i متن iام مجموعه آموزشی و w_k^i کلمه kام در متن iام بوده و تابع $\#(.)$ تعداد آیتم‌ها است. حال با توجه به تئوری بیز، احتمال تعلق هر متن به هر دسته طبق (2-3) محاسبه می‌گردد.

$$p(c_j|d_i) = \frac{p(c_j) * p(d_i|c_i)}{\sum_{c_j \in C} p(c_j) * p(d_i|c_i)} \qquad (3-2)$$

از مهم‌ترین ویژگی‌های این روش سادگی آن و مقاوم‌بودن در برابر خصیصه‌های نامرتبط است. این روش مبتنی بر قانون احتمالات بوده و نسبت به شبکه عصبی و درخت تصمیم‌گیری کارایی بالایی دارد، دقت و سرعت بالا در مجموعه داده‌های بزرگ، مستقل بودن از خصیصه‌ها، عدم تحمل و انطباق-پذیری با داده حقیقی و بلادرنگ نیز از دیگر خصیصه‌های آن محسوب می‌شود.

در بسیاری از موارد از مدل توسعه یافته این روش موسوم به n- گرام‌ها استفاده می‌گردد. این روش برخلاف بیز محل قرارگیری کلمات نیز در احتمال رخداد متن اهمیت دارد[21,22,23] .

روش k- نزدیک ترین همسایه

در این روش به جای استفاده از تمام اسناد، از k سند در نزدیک‌ترین همسایگی مورد نظر استفاده می‌کند. در این متد از اندازه کسینوس میان دو بردار سند برای تعیین میزان شباهت استفاده می‌شود. در این روش، همسایه‌های نزدیک انتخاب شده و به وسیله امتیازدهی دسته همسایه برنده به سند جدید اختصاص می‌یابد. در صورتی که سند به بیش از یک همسایه شباهت داشته باشد، از مجموع امتیازها به عنوان وزن نهایی استفاده می‌کنیم. مهم‌ترین مشکل در این روش تعیین تعداد همسایه مجاور برای پردازش و حد آستانه‌ای برای انتساب سند به گروه‌ها است[23].

روش ماشین بردار پشتیبان[42]

این الگوریتم که به روش SVM معروف است، مبتنی بر روش‌های آماری بوده و اطلاعات را از فضای برداری حاضر به فضای برداری دیگری (عموما در ابعاد بالاتر) می‌برد. این عمل توسط توابع کرنل صورت گرفته و پس از نگاشت، حداکثر حاشیه ممکن میان نمونه‌های مثبت و منفی از طریق بهینه-سازی مرز داده‌ها انجام می‌گیرد. از مزایای این روش می‌توان به عدم وابستگی آن به تعداد نمونه‌های آموزشی و کارایی بالا در تعداد مشخصه زیاد و نمونه کم نام برد. بسته به فضای کار، داده‌ها توسط

[42] Support Vector Machine

خطوط و یا صفحات از هم جدا می‌شوند. این الگوریتم در حال حاضر کاربرد بسیار بالایی در اکثر موارد و خصوصا در متن‌کاوی پیدا کرده است[24,25].

3-3-2 دسته‌بندی اسناد متنی

دسته‌بندی متن به معنای انتساب اسناد متنی بر اساس محتوای آنها به یک یا چند طبقه از قبل تعیین شده است که یکی از مهمترین مسائل حوزه متن‌کاوی می‌باشد. متن‌کاوی، فرایند کشف نیمه اتوماتیک الگوها و مسیرها در مجموعه عظیمی از متن‌های غیرساخت‌یافته است. متن‌کاوی به عنوان روشی در استخراج دانش از متون، یکی از موضوعات مهم در گستره ای از اعمال مدیریت اطلاعات است.

یکی از روش‌های مورد استفاده در متن‌کاوی، استفاده از تکنیک‌های یادگیری ماشین است. مسأله قابل تأمل این است که این تکنیک‌ها ابتدا در مورد داده‌های ساخت‌یافته به کار گرفته شدند و علمی به نام داده‌کاوی را به وجود آوردند. داده‌های ساخت‌یافته به داده‌هایی اطلاق می‌گردد که به طور کاملا مستقل از یکدیگر ولی یکسان از لحاظ ساختاری در یک محل گردآوری شده‌اند[26].

بنابراین تفاوت بارز داده‌کاوی و متن‌کاوی در این است که تمرکز اصلی داده‌کاوی بر داده‌های ساخت یافته و تمرکز متن‌کاوی بر داده‌های نیمه ساخت‌یافته و غیرساخت‌یافته می‌باشد .داده‌کاوی تلاش می‌کند تا از میان مجموعه داده‌های فراوان، الگوها و روابطی را در اختیار تحلیلگرها و تصمیم‌گیرنده‌ها قرار دهد[27]. به هر حال استفاده از داده کاوی در مورد متن، بنیان‌گذار شاخه‌ای جدید در علوم هوش مصنوعی گردید که امروزه متن‌کاوی نامیده می‌شود. در کاربرد خوشه‌بندی متن، هر سند به عنوان یک داده آموزشی در نظر گرفته می‌شود .هر خوشه در این کاربرد، حاوی اسناد مربوط به یک موضوع خاص است .تعداد خوشه‌های آموزشی بر اساس تنوع موضوعی و توسط فردی که داده‌های آموزشی را فراهم می‌کند، مشخص می‌شود .برچسب داده‌های آموزشی، شماره و یا نام خوشه و یا خوشه‌هایی است که هر داده آموزشی به آن خوشه و یا خوشه‌ها تعلق دارد. عامل یادگیر در فاز آموزشی با استفاده از داده‌های آموزشی(داده های برچسب دار)، آموزش داده شده و یک تعمیم از هر

خوشه، عامل را قادر به شناسایی اسناد مربوط به هر خوشه می‌کند .در فاز آزمایش، عامل یادگیر، اسناد متنی را که در فاز آموزشی ندیده است و متعلق به یکی از خوشه های آموزشی می‌باشد، دریافت کرده و احتمال تعلق این سند را به هر کدام از خوشه‌های آموزشی به دست می‌آورد. در این-جا خوشه بندی برای هر سند، به معنی انتخاب خوشه‌ای با بالاترین احتمال تعلق سند به آن خوشه می‌باشد.

2-3-3-1 آماده سازی و پیش پردازش داده‌ها

داده‌های خام جمع‌آوری شده معمولا دارای حجم زیاد و بسیار ناهمگن می‌باشند. این داده ها باید به داده‌های سازگار و یکپارچه تبدیل شوند تا بتوانند برای مرحله‌ی کشف الگو مفید واقع باشند. به فرایند پردازش داده‌های خام به منظور تبدیل به فرمت مناسب برای آموزش، پیش پردازش گفته می-شود. همانند بیشتر کاربردهای داده‌کاوی، پیش‌پردازش و آماده‌سازی داده شامل پاکسازی و حذف نویز، تبدیل فرمت داده، یکپارچه‌سازی و همچنین برطرف کردن ناسازگاری‌ها می‌باشد[28,29,30]. کارهای عمده‌ای که اغلب در آماده‌سازی یا پیش‌پردازش داده‌ها صورت می‌گیرد، از این قرارند:

ریشه یابی

عمل ریشه یابی [31,32,33,34] با هدف زدودن الحاقات و یافتن جوهره اصلی واژه به کمک جداول راهنما و حذف پیشوند و پسوندهای مرسوم در هر زبان انجام می‌گیرد. ریشه‌یابی لغات به معنای زبان‌شناسی آن نبوده بلکه صرفا به معنای دسته‌بندی کلمات در گروه‌های معنایی یکسان است. بدین صورت شیوه نوشتار تمام کلمات هم‌ریشه یکسان خواهد گشت. برای این کار نیاز به شناخت کافی از دستور زبان داشته و باید ماشین حالت مناسب و متناسب با ساختار زبان تهیه نماییم. الگوریتم‌های موجود در این زمینه مبتنی بر دیکشنری و قانون بوده که مورد اول به دلیل عدم قابلیت گسترش‌پذیری، کارایی بالایی ندارد. معروف‌ترین ریشه‌یاب در زبان انگلیسی Porter [35] بوده و تاکنون چندین نمونه فارسی بر اساس آن نوشته شده است. در زیر به بیان کلی تفاوت‌های میان این دو دسته می‌پردازیم.

هر دو ریشه یاب به دنبال پسوندهای خاصی جستجو می‌کنند و مراحل مختلفی را بر طبق لیست قوانین پسوندی پشته‌گذاری شده طی می‌کنند. . با این حال تفاوت‌های مهمی بین این دو الگوریتم وجود دارد. برای مثال الگوریتم ریشه‌یاب پورتر به منظور تخمین محتوای اطلاعات، الگوی حروف صدادار و بی‌صدا را تشخیص می‌دهد؛ اما در فارسی بسیاری از حروف صدادار نوشته نمی‌شوند، بنابراین ریشه‌یاب فارسی از طول رشته برای تعریف کران پائین محتوای ریشه استفاده می‌کند.(در حال حاضر مینیمم طول ریشه 3 است). این محدودیت در بعضی موارد باعث خطا می‌گردد بخصوص زمانی‌که یک زیررشته که قسمتی از یک کلمه کوتاه است ، به اشتباه به عنوان یک پسوند در نظر گرفته شود. تفاوت دیگر این دو الگوریتم آن است که این الگوریتم بر خلاف الگوریتم پورتر ، پیشوندها را هم شناسایی می‌کند.

حذف کلمات عمومی

در این مرحله به تشخیص، علامت گذاری و حذف کلمات عمومی [36,37,38] می‌پردازیم. به صورت کلی می‌توان کلمات موجود در یک متن را به سه دسته عمومی، خاص و کلیدی تقسیم نمود. کلمه عمومی، کلمه‌ای است که کلیه اسناد تمام کلاس ها بارها و بارها ظاهر شده و اهمیت و مفهوم چندانی ندارند و معمولا در فاز پیش پردازش ، این کلمات حذف می‌شوند. به عنوان نمونه می‌توان از ضمایر، قیود، حروف ربط و اضافه، افعال و ... نام برد. این کلمات در محتوای سند هیچ‌گونه ارزشی ندارند. کلمات خاص، کلماتی با فراوانی زیاد فقط در کلیه اسناد یک کلاس و فراوانی کم در سایر کلاس‌ها بوده و هم‌پوشانی حداقلی آنها، قابلیت مجزا سازی بهینه کلاس ها را به میزان زیادی افزایش می‌دهد. در نهایت کلمات کلیدی به عبارات یا کلمات مهم برای تشریح محتویات داخل سند گفته شده و پس از وزن‌دهی به کلمات کاندید اولیه انتخاب می‌گردند. این گروه دارای فراوانی زیاد در سند خاص بوده و در ایجاد خلاصه متن اهمیت بسیار بالایی دارند. در اکثر زبان‌ها لیستی از کلمات عمومی پرتکرار وجود دارد، که در مرحله پیش پردازش آنها را حذف می‌نماییم. عملیات حذف در دو مرحله قبل و بعد از عمل ریشه‌یابی صورت گرفته و نقش بسزایی در کاهش تعداد کلمات قابل پردازش دارد.

تقسیم‌بندی

هدف از این کار تشخیص محدوده جملات، عبارات و یا لغات قابل پردازش در یک متن است. به عنوان مثال در ساده‌ترین روش می‌توان از کاراکترهای خالی و علائم و نشانه‌های نقطه‌گذاری در متن استفاده نمود. در زبان فارسی با مدل‌های مختلف فضای خالی کامل و نصفه مواجه هستیم که به ترتیب برای جداسازی کلمات تک واژه‌ای و کلمات چند واژه‌ای به کار برده می‌شوند. هرچند در اکثر موارد هنوز استفاده از نیم فاصله مرسوم نبوده و برای بالا بردن دقت باید جدولی شامل کلمات چند بخشی ایجاد گردد.

لازم به ذکر است که تمامی موارد مورد بحث در قسمت پیش‌پردازش در صورتی کارایی مورد نظر را خواهند داشت که مطابق با زبان سیستم طراحی شده باشند، چرا که ویژگی‌ها و قوانین زبان‌شناسی، دستوری و نگارشی برای هر زبان متفاوت بوده و از استاندارد مشخصی پیروی نمی‌کنند.

2-3-3-2 وزن دهی به کلمات و ایجاد ماتریس

وزن‌دهی ویژگی به عنوان معیاری عددی برای محاسبه میزان تاثیر ویژگی خاص در مجموعه می‌باشد. به عنوان مثال اگر مجموعه کلی ما دارای i سند مجزا و k ویژگی متمایز کننده باشد، در نتیجه برای هر سند، بردار وزن $(1,k)$ مانند رابطه زیر ایجاد خواهد گشت.

$$\text{اگر } d_i \in D \rightarrow d_i = (w_{1i}, \dots, w_{ki})$$

هدف از وزن‌دهی به کلمات تعیین ارزش و اهمیت آنها در یک سند است. در بسیاری از موارد کلمات کلیدی نهایی بر اساس وزن به دست آمده از میان کلمات کاندید انتخاب می‌شوند. در این زمینه تا-کنون روش‌های بسیاری به کار برده شده است که اکثرا نتیجه ترکیب یک یا چند مورد از موارد زیر است که بسته به کاربرد و نوع داده ورودی از آن استفاده نموده‌اند. در ساده‌ترین شکل وزن‌دهی به صورت منطق (1/0) عمل کرده و وجود کلمه در سند را با میزان عددی "یک" و عدم وجود آن را با مقدار "صفر" نمایش می‌دهیم [39]. در نهایت پس از وزن‌دهی به عبارات و لغات و تعیین ارزش و نقش آنها در اسناد، داده‌ها را به فضای محاسباتی برده و برای این منظور ماتریسی شامل تمام اسناد و

لغات تعیین شده رسم می‌کنیم. به گونه‌ای که هر سطر نشان‌دهنده یک متن، هر ستون بیان‌کننده یک لغت و سلول متناظر با هر سطر و ستون نشان‌دهنده وزن لغت در سطر باشد. بدیهی است عدم وجود عبارت در یک سند معادل با مقدار صفر خواهد بود. نمایش ماتریس به صورت شکل (2-2) است.

$$D = \{d_1, d_2, ..., d_n\}$$
$$C = \{c_1, c_2, ..., c_m\}$$
$$a_{ij} = \begin{cases} w_{ij} & d_j \in c_i \\ 0 & d_j \notin c_i \end{cases}$$

شکل(2-2): ماتریس سند-لغت، نمایانگر وزن هر لغت در سند متناظر

به صورت کلی روش‌های متداول وزن‌دهی ویژگی را بر حسب توابع توزیع آنها می‌توان به سه گروه مختلف دسته‌بندی نمود. در ادامه لیستی از مهم‌ترین این روش‌ها آورده شده است.

فرکانس تکرار

اساس این روش توجه به کلمات یا عبارات کاندید به عنوان موجودیتی مستقل بدون توجه به بار معنایی و نحوه قرارگیری در متن است. بدین منظور که صرف وجود یا روئیت کلمه در متن، به آن اعتبار خواهد داد.

فرکانس تکرار لغت در سند[43]

این متد به صورت محلی به اهمیت و میزان تکرار لغت یا عبارت در یک سند مجزا می‌پردازد. انواع مختلف وزن‌دهی بر اساس فرکانس لغت در سند در جدول (2-4) آورده شده است. در تمام موارد t_k بیانگر kامین واژه یا عبارت و d_i نشان‌دهنده iامین سند قابل پردازش است[40,41] .

[43] Term frequency

نام روش	رابطه	توضیحات
TF	$w_{ki} = tf(t_k, d_i) = \begin{cases} \#(t_k, d_i) & t_k \in \text{vector of } d_i \\ 0 & t_k \notin \text{vector of } d_i \end{cases}$	$\#(t_k, d_i)$ تعداد تکرار ویژگی t_k در مستند d_i
normTF	$w_{ki} = \frac{tf(t_k, d_i)}{\sqrt{\sum_k (tf(t_k, d_i))^2}}$	فرمول نرمال شده
logTF	$w_{ki} = \log(tf(t_k, d_i))$	لگاریتم فرمول
ITF	$w_{ki} = ITF(t_k, d_i) = 1 - \frac{1}{1 + tf(t_k, d_i)}$	r=1
Sparck	$w_{ki} = Sparck(t_k, d_i) = tf(t_k, d_i) * (k - \log(p_k))$ $p_k = \sum_D tf(t_k, d_i)$	k تعداد ویژگی های منحصر به فرد در مجموعهٔ D

جدول(2-4): روش‌های وزندهی بر اساس فرکانس تکرار لغت در سند

معکوس فرکانس تکرار لغت در اسناد[44]

این شیوه به صورت عمومی عمل کرده و هدف اصلی آن تلاش برای کاهش وزن ویژگی با افزایش فرکانس ویژگی در مجموعه مستندات است. چرا که هر چه کلمه‌ای بیشتر در کل متون تکرار شده باشد، جنبه عمومی‌تری گرفته و برای جداسازی حداکثری دسته‌ها از هم قابلیت کمتری خواهد داشت. در اکثر موارد از ترکیب این روش و روش فوق استفاده می‌گردد. جدول (2-5) فرمول اصلی و موارد ترکیبی را نشان می‌دهد. در تمام موارد t_k بیانگر kامین واژه یا عبارت و d_i نشان‌دهنده iامین سند قابل‌پردازش است[41,42].

[44] *Inverse document frequency*

نام روش	رابطه	توضیحات
IDF	$w_{ki} = idf(t_k, d_i) = \log \frac{\mid D \mid}{\mid D(t_k) \mid}$	|D|: تعداد کل اسناد مجموعه D |D(t_k)|: تعداد مستندات شامل ویژگی t_k
TFIDF	$w_{ki} = tfidf(t_k, d_i) = tf(t_k, d_i) * idf(t_k, d_i)$	ترکیبی از فرکانس تکرار محلی و عمومی
normTFIDF	$w_{ki} = \frac{tfidf(t_k, d_i)}{\sqrt{\sum_k (tfidf(t_k, d_i))^2}}$	فرم نرمال شده

جدول(2-5): روش‌های وزندهی بر اساس معکوس فرکانس تکرار لغت در اسناد و ترکیب آن با روش محلی

فرکانس تکرار در دسته[45]

همان‌گونه که پیشتر بیان شد هدف اصلی در فرآیند دسته‌بندی، ایجاد گروه‌ها و دسته‌های متمایز از یکدیگر است. هر چه تفاوت میان دسته‌ها بیشتر باشد، انتساب اسناد به گروه‌های مربوطه به صورت دقیق‌تری صورت می‌گیرد. در نتیجه باید تا حدامکان ارزش کلمات پرکاربرد عمومی کاهش یابد تا کلمات کلیدی مهم‌تر و خاص‌تر استخراج گردند. در تمام موارد t_k بیانگر kامین واژه یا عبارت، d_i نشان دهنده iامین سند و c_j بیانگر jامین دسته قابل پردازش است[41,43]. دو فرمول مهم در این زمینه در جدول (2-6) آورده شده است.

[45] Cluster frequency

نام روش	رابطه	توضیحات
TFRF	$w_{ki} = TFRF(t_k, d_i) = \frac{tf(t_k, d_i) * rf(t_k, c_{d_i})}{\sqrt{\sum_k (tf(t_k, d_i))^2 * (rf(t_k, c_{d_i}))^2}}$ $rf(t_k, c_j) = \log\left(2 + \frac{\lvert D(t_k, c_j)\rvert}{\sum_{m=1, m\neq j}^{\lvert C\rvert} \lvert D(t_k, c_m)\rvert}\right)$	$\lvert D(t_k, c_j)\rvert$: تعداد مستندات مجموعه D و طبقه c_j شامل ویژگی t_k $\sum_{m=1, m\neq j}^{\lvert C\rvert} \lvert D(t_k, c_m)\rvert$: مجموع تعداد مستندات D و طبقه غیر از c_j شامل ویژگی t_k
TFCRF	$w_{ki} = TFCRF(t_k, d_i) = \frac{\log(tf(t_k, d_i) * crfValue(t_k, c_{d_i}))}{\sqrt{\sum_k (\log(tf(t_k, d_i) * crfValue(t_k, c_{d_i})))^2}}$ $negativeRF(t_k, c_i) = \frac{\sum_{m=1, m\neq j}^{\lvert C\rvert} \lvert D(t_k, c_m)\rvert}{\sum_{m=1, m\neq j}^{\lvert C\rvert} \lvert D(c_m)\rvert}$ $positiveRF(t_k, c_i) = \frac{\lvert D(t_k, c_j)\rvert}{\lvert D(c_j)\rvert}$ $crfValue(t_k, c_i) = \frac{positiveRF(t_k, cj)}{negativeRF(t_k, c_j)}$	$\lvert D(c_j)\rvert$: تعداد مستندات طبقه c_j

جدول(2-6): روش‌های وزن‌دهی بر اساس فرکانس تکرار لغت در دسته

مجاورت معنایی[46]

در این روش به کمک مفاهیم آنتولوژی و زبان‌شناسی کلمات مترادف با کلمه کلیدی تطابقت داده شده و بدین وسیله امتیاز کلمه در سند افزایش خواهد یافت. بدین منظور باید از دیکشنری‌های لغت برای یافتن کلمات هم‌معنی و افزودن آنها به متن استفاده نماییم. از مهم‌ترین دیکشنری‌ها در این زمینه می‌توان به wordnet و farsnet اشاره نمود[44,45].

موقعیت کلمات[47]

یکی دیگر از عوامل موثر در وزن‌دهی به کلمات، موقعیت مکانی آنها خواه به صورت مستقل و خواه بر حسب دیگر لغات کلیدی است. روش‌های بسیار زیادی در این قسمت نیز وجود دارد که بسته به کاربرد با افزایش ضریب خاص، ارزش کلمه را زیاد می‌کنند. علاوه بر موقعیت, نوع نگارشی (حروف بزرگ/ کوچک)، نوع فونت (افزایش)، نوع فرمت (ضخیم/کج) و همچنین طول عبارت (شامل چندین

[46] Semantic proximity
[47] Term position

کلمه/ طول یک کلمه) نیز در وزن‌دهی اهمیت بسیار بالایی دارد. به عنوان مثال قرار گرفتن کلمه در عنوان مطلب یا فرمت نوشتاری خاص وزن بیشتری به کلمات اختصاص خواهد داد. همچنین اسنادی که در آنها کلمات کلیدی به یکدیگر نزدیک‌تر هستند، به احتمال زیاد اسناد مشابه‌تری برای بازیابی می‌باشند.

2-3-3-3 کاهش داده و کاهش بعد

کار با اسناد متنی به دلیل تنوع و حجم بالای اطلاعات اغلب زمانبر است. هدف از این مرحله، دست یافتن به حجم کوچکتری از داده‌ها جهت بالا بردن زمان پردازش و کاهش حافظه مصرفی است. نکته مهم در این مرحله از آماده‌سازی داده‌ها آن است که دستیابی به نتایج تحلیلی مشابه با اصل و تمام داده‌ها تضمین گردد، چرا که در غیر این صورت این کاهش، اثر مثبتی برای ما در پی نخواهد داشت. هر چند در بسیاری از موارد از این تکنیک‌ها صرف‌نظر شده و تنها سعی در یافتن و بهینه‌سازی کلمات کلیدی حداقلی است.

2-3-3-4 معیارهای شباهت

معیارهای سنجش شباهت، ابزاری ضروری در بسیاری از کاربردها مانند مسائل تصمیم‌گیری، کلاس-بندی الگوها وداده‌کاوی هستند. بیشترین کاربرد در این زمینه مباحث مربوط به کلاس‌بندی داده‌ها برای یافتن کلاس مربوط به یک داده جدید و وارد کردن آن به کلاس در مسائل طبقه‌بندی است. از این‌رو می‌توان بیان داشت که معیار شباهت نقش مهمی را در پیدا کردن ارتباطات میان داده‌ها و در نتیجه میزان عملکرد سیستم دارد. انتخاب یک معیار شباهت درست و کارا بستگی مستقیم به نوع داده ورودی داشته و باید در هر مسئله از روش مناسب استفاده نمود. وابسته به کاربرد مورد نظر ممکن است یک معیار شباهت k بعدی برای سنجش شباهت مورد نیاز باشد .در حال حاضر معیارهای شباهت بسیاری برای اندازه گیزی میزان شباهت وجود دارند.این معیارها در سه گروه مختلف طبقه

بندی شده‌اند: میزان فاصله[48] ، میزان نزدیکی[49] و میزان باینری[50]. مهم ترین معیارها به همراه فرمول-ها و پارامترهای مربوط به دو گروه اول به ترتیب در جداول (2-7) و (2-8) آورده شده است.

در معیارهای مبتنی بر فاصله ، معیارها ومشخصه‌های اختصاصی یک داده متقابلا با مشخصه‌های یک داده دیگر مورد مقایسه قرار می‌گیرد.به عبارت دیگر در این گروه فقط داده‌های اولیه وناخالص دو مجموعه به منظور تشخیص میزان شباهت مورد بررسی قرار می‌گیرند .در معیارهای مبتنی بر نزدیکی اطلاعات اضافه دیگری در مورد داده‌ها مانند مقدار میانگین ،انحراف استاندارد نیز برای سنجش میزان شباهت استفاده می‌شوند .معیارهای باینری نیز برای اندازه‌گیری میزان شباهت در داده‌های باینری مورد استفاده قرار می‌گیرند . در تمام موارد y , x به دو سند متفاوت از مجموعه اسناد که قصد بررسی میزان تشابه میان آنها را داریم، اشاره دارند[46].

Row num.	Distance Measure Methods	Calculation Formula
1	Euclidian	$Euclid(x,y) = \sqrt{\sum(x_i - y_i)^2}$
2	Chebychev	$Chebychev(x,y) = max_i\lvert x_i - y_i\rvert$
3	Block(Hamming)	$Block(x,y) = \sum\lvert x_i - y_i\rvert$

جدول(2-7): روش‌های اندازه‌گیری معیار شباهت بر مبنای فاصله

[48] Distance Measures
[49] proximity measures
[50] binary measures

Row num.	Proximity Measure Methods	Calculation Formula
1	Correlation	$Correlation(x,y) = \frac{\sum(Z_{xi}Z_{yi})}{N}$
2	Cosine	$Cosine(x,y) = \frac{\sum(x_i y_i)}{\sqrt{((\sum x_i^2)(\sum y_i^2))}}$
3	Chi-Square	$Chisq(x,y) = \sqrt{\sum \frac{(x_i - E(x_i))^2}{E(x_i)} + \sum \frac{(y_i - E(y_i))^2}{E(y_i)}}$
4	Jensen	$Jensen(x,y) = \frac{1}{2}\sum_{i=1}^{k}\{x_i` log_2 x_i` + y_i` log_2 y_i` - (x_i` + y_i`) log_2(\frac{x_i` + y_i`}{2})\}$ $where \quad x_i` = \frac{x_i}{\sum_{i=1}^{k} x_i} \quad , \quad y_i` = \frac{y_i}{\sum_{i=1}^{k} y_i}$

جدول (2-8): روش‌های اندازه‌گیری معیار شباهت بر مبنای نزدیکی

در انتها عمل دسته بندی بر اساس یکی از روش‌های مطرح شده در بخش مربوطه و به کمک یکی از معیارهای اندازه‌گیری بسته به نوع و ذات‌داده های ورودی انجام می‌پذیرد.

2-4 دسته‌بندی اسناد وب

با افزایش رو به رشد اطلاعات در بستر وب، یافتن سریع اطلاعات مورد نظر درخواستی از موتورهای جستجو دشوارتر گشته است. یکی از راهکارهای حل این معضل استفاده از تکنیک‌های دسته‌بندی اسناد و گروه‌بندی آنها در رده‌های مشابه به منزله تسهیل در نمایش و ارائه نتایج به صورت فشرده است. دسته‌بندی به معنای روند شکل‌دهی گروه‌هایی از اشیاء مشابه از یک مجموعه داده ورودی است. به عنوان یکی از مشخصه‌های دسته‌بندی خوب باید تمام عناصر درون‌گروهی با یکدیگر بیشترین میزان شباهت را داشته و با عناصر دیگر گروه‌ها دارای کمترین وجه تشابه باشند. هر چند ایده اصلی دسته‌بندی از محاسبات آماری روی داده‌های عددی سرچشمه گرفته است، اما نیازهای موجود در علوم کامپیوتر و به طور خاص داده‌کاوی، این مفهوم را در دیگر حوزه‌ها مانند داده‌های متنی و چند

رسانه‌ای نیز گسترش داده است. دسته‌بندی اسناد وب در ابتدا در سایت‌هایی مانند dmoz.org و یاهو تا سال 1998 به صورت دستی به کمک افراد خبره و کارشناسان صورت می‌گرفت. هرچند این نوع دسته‌بندی دارای درجه بالا دقت بود، اما ورود تمایلات و تصمیمات شخصی و عدم تشابه آن میان افراد مختلف و همچنین حجم بالا اطلاعات از کارایی سیستم‌ها کاسته و متعاقبا نیاز به تعداد نیروی انسانی زیاد و صرف زمان و هزینه بالا وجود داشت. لذا این وظیفه بر عهده ماشین‌های اتوماتیک گذاشته شد.

در مبحث دسته‌بندی نتایج جستجو وب سعی در بسط این ایده در اسناد بازگردانده شده از موتورهای جستجو (برحسب درخواست یا پیش فرض) داریم، به گونه‌ای که اسناد دسته‌بندی شده به کاربر در نحوه و زمان دسترسی به اطلاعات مورد نیاز و علاقه خود کمک رسانده و عمل دسترسی به اطلاعات را تسهیل بخشد. در این سیستم هر کدام از قطعات[51] خبری را به یک گروه موضوعی معنادار نسبت می‌دهیم. این اقدام به معنای پس پردازش نتایج نهایی به دست آمده از موتورهای جستجوگر به منزله نمایش بهینه و سهولت دسترسی به اطلاعات است. چرا که بدون وجود موتورهای جستجوگر، اینترنت به حجم انبوهی از اطلاعات آشفته و بدون سازمان تبدیل می‌شود که عملا استفاده از آن برای هیچ کس مقدور نخواهد بود. بر خلاف باور عمومی مردم موتورهای جستجوگر به هیچ عنوان جواب سوالات را ندارند، بلکه صرفا راهی برای دسترسی سریع به جواب و اطلاعاتی که مردم دیگر بر روی وب قرار داده‌اند را پیش‌روی کاربر قرار می‌دهند و اغلب به جای پاسخ دادن، تنها صفحات متناظر با درخواست کاربر را بازمی‌گردانند و رضایت کاربر را در یافتن اطلاعات مورد نظر بدست می آورند. مشکل اصلی زمانی رخ می دهد که موضوع درخواست عبارتی مبهم، دارای گستردگی مفهومی زیاد و یا عبارتی ضعیف از لحاظ معنایی باشد. در این هنگام با طیف وسیعی از موضوعات نامربوط، گیج کننده و بی-ارزش برای کاربر مواجه خواهیم شد. هدف موتور دسته‌بند مطالب، ارائه بهتر و فشرده‌تر موارد مشابه با هم به کاربر است.

[51] Snippet

چهار معیار اساسی در ایجاد دسته‌ها عبارتند از: ایجاد عناوین مختصر، دقیق و مناسب، مشخص و مورد علاقه کاربر انسانی است، به طوری که مشخص نشود این عبارت توسط یک ماشین ایجاد شده است. یکی از مشخصه‌های موتورهای دسته‌بند اسناد، عدم حفظ عنوان و یا خود داده‌ها است، بلکه در اکثر موارد از نتایج بدست آمده از موتورهای جستجوگر دیگر استفاده می‌کنند. دقت پایین موتورها و لیست‌های طولانی نمایش نتایج کار یافتن اطلاعات مورد نظر را برای کاربران بسیار دشوار و زمانبرکرده است. در مواقعی حتی درخواست مشابه در چند موتور نتایج متفاوتی ارائه می‌دهد که بیشتر آنها نسبتی با مورد درخواستی ندارند. در نتیجه استفاده از موتور دسته بند قطعات وب، برای ایجاد مکملی در نمایش لیست نتایج مفید به نظر می‌رسد. این موتورها به نام موتور دسته بند نتایج وب نیز معروف هستند. در این روش قطعات خبری بدست آمده از موتورهای جستجوگر در گروه‌های موضوعی معنادار دسته بندی می‌شوند. این مقوله در هر دو زمینه تجاری و علمی تاکنون پیشرفت‌های بسیاری داشته است.

هرچند دسته‌بندی نتایج وب وابستگی بسیار شدیدی به دسته‌بندی اسناد دارد ، اما چالش‌های بسیاری در این زمینه دیده می‌شود که اثر بخشی و بهره‌وری الگوریتم‌های پایه دسته‌بندی اسناد را تحت تاثیر قرار می‌دهد. اولین تفاوت تاکید بر روی کیفیت برچسب خوشه‌ها است تا جایی که مقالات بسیاری توجه خود را بیشتر بر روی این موضوع معطوف کرده اند.

دسته بندی متون عمدتا به عنوان روشی برای افزایش اثربخشی مرتب سازی اسناد به کار می‌رود، چرا که به صورت منطقی و واضح، اسناد مشابه به شدت مربوط به درخواست‌های مشابه می‌باشند. این دسته‌بندی پایه و مبنای بسیاری از اقدامات از جمله دسته بندی اسناد وب و طراحی موتورهای دسته بند است. الگوریتم‌های دسته‌بندی نتایج وب بنا بر نظر [18] باید دارای ویژگی‌های زیر باشند.

1: ارتباط داشتن[52]: روش باید دسته‌هایی ایجاد نماید که قادر به گروه‌بندی اسناد مربوطه با درخواست کاربر به صورت مجزا از اسناد غیر مربوطه ایجاد نماید.

[52] Relevance

2: قابلیت مرور خلاصه[53]: کاربر باید در یک نگاه کوتاه قادر به تشخیص مفید بودن یا نبودن دسته برای کارایی موردنظر باشد. هدف تنها جایگزینی دسته‌ها با نتایج مرتب شده نیست، در نتیجه الگوریتم باید قادر به ارائه توضیح مختصر و مفید از خوشه باشد.

3: هم پوشانی[54]: از آنجا که هر سند دارای چندین موضوع است، نباید هیچ محدودیتی در اختصاص اسناد به چندین گروه وجود داشته باشد.

4: میزان تحمل پذیری در استفاده از قطعات[55]: الگوریتم باید قادر به ایجاد دسته‌هایی با کیفیت بالا باشد، حتی اگر برای این کار از قطعات خبری برگرفته از موتورهای جستجوگر استفاده شود. چرا که در اکثر موارد کاربر تحمل انتظار برای بارگزاری کامل متن را ندارد.

5: سرعت[56]: کاربران بسیاری وجود دارند که بنا به حساسیت بر روی جستجو، تمایل بررسی بیش از 100 قطعه اولیه نتایج را دارند، یکی از دلایل دسته بندی، اجازه دادن به کاربر برای مطالعه موضوعات مشخص و در نتیجه صرفه جویی در زمان وی است، در نتیجه الگوریتم دسته‌یابی باید قادر به گروه-بندی هزاران قطعه در کمترین زمان ممکن باشد. برای کاربران حساس، حتی ثانیه‌ها نیز ارزشمند هستند.

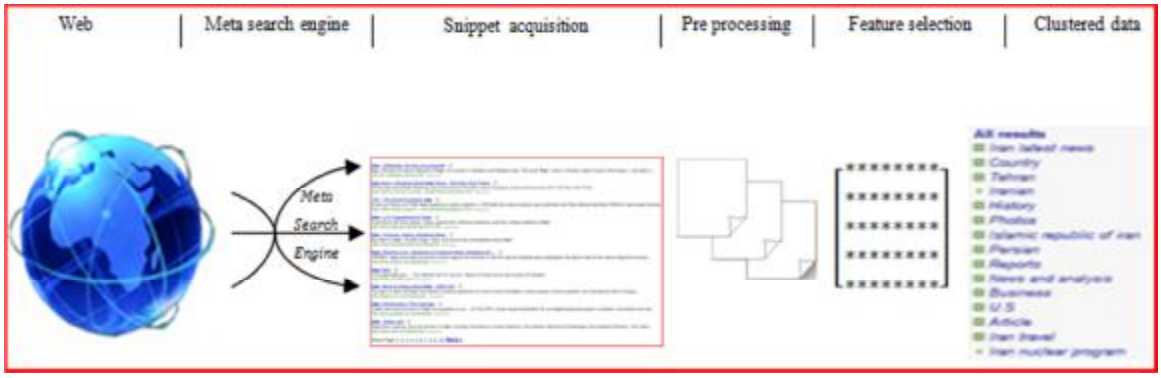
6: قابلیت افزایشی[57]: به منزله صرفه جویی در زمان روش مورد نظر باید هر قطعه را به محض ورود به وب مورد پردازش قرار دهد.

2-4-1 مراحل خوشه بندی نتایج وب

مراحل خوشه بندی نتایج وب در شکل (2-3) آورده شده و در ادامه توضیح هر کدام با جزئیات بیشتر بیان گشته است. در این قسمت از دوباره‌گویی مطالب مشترک با تکنیک‌های پردازش متن چشم-پوشی کرده و تنها در مواردی که تفاوت خاصی وجود دارد، به آنها اشاره نموده‌ایم.

---

[٥٣] Browsable summaries
[٥٤] Overlap
[٥٥] Snippet- Tolerance
[٥٦] Speed
[٥٧] Incrementality



شکل(2-3): مراحل خوشه‌بندی نتایج وب

وب: در حال حاضر وب به عنوان بزرگترین منبع اطلاعات درهم و غیر ساخت‌یافته شناخته شده است. برای بدست‌آورن اطلاعات، تعدادی عامل توسط موتورهای جستجوگر در صفحات پخش شده و اطلاعات مشابه با درخواست مطرح شده و یا اطلاعات خام مورد نیاز را بر می‌گردانند.

موتورهای جستجوگر چندگانه: به دلیل عدم پوشانندگی کامل وب توسط یک موتور جستجوگر منفرد، برای دستیابی کامل و همه جانبه به اطاعات پیرامون یک موضوع خاص نیاز به ادغام چندین موتور وجود دارد. در همین راستا از موتورهای جستجوگر چندگانه استفاده می‌نماییم که به جای پویش و ایندکس‌گذاری صفحات وب و جمع‌آوری و ذخیره نتایج در یک پایگاه، از نتایج بدست‌آمده از دیگر موتورها استفاده می‌نمایند. یکی از اقدامات مهم حذف منطقه همپوشانی موتورهای جستجو است بدین صورت که عناوین عینا مشابه که از سایت منبع و زمان یکسان برگردانده شده از میان لیست حذف شده و تنها قطعه مربوط به یکی از موتورها باقی می‌ماند.

تهیه قطعات: هدف این فاز جمع‌آوری قطعات متنی است که توسط موتورهای جستجو در جواب درخواست کاربر بازگردانده شده‌اند. تمام اطلاعات مورد نیاز در این قسمت از صفحات وب استخراج شده و تگ‌های اضافه و HTML آن جدا می‌شوند. در این قسمت تنها عنوان، محتوای قطعه ، منبع اولیه و تاریخ ثبت نگهداری می‌شود. در این قسمت صفحه‌ای مانند صفحه نتیجه جستجو گوگل با لیستی ترتیبی از اطلاعات مشابه خواهیم داشت. این اطلاعات می‌توانند بر حسب درخواست کاربر و یا

بر حسب پیش‌فرض سیستم بازگردانده شوند که در مورد اول باید به پیش پردازش درخواست کاربر و اعمال قوانین جبری منطقی احتمالی تعبیه شده در صفحه رابط کاربر پرداخته شود.

پیش‌پردازش: عمل پیش‌پردازش به منزله یکدست‌سازی اطلاعات، حذف نویز و آماده‌سازی نهایی داده‌ها برای پردازش انجام می‌گیرد. در این قسمت عملیات فیلتر سازی، حذف کلمات عمومی، ریشه-یابی و تقسیم نمودن سند به بخش‌های قابل پردازش انجام می‌پذیرد.

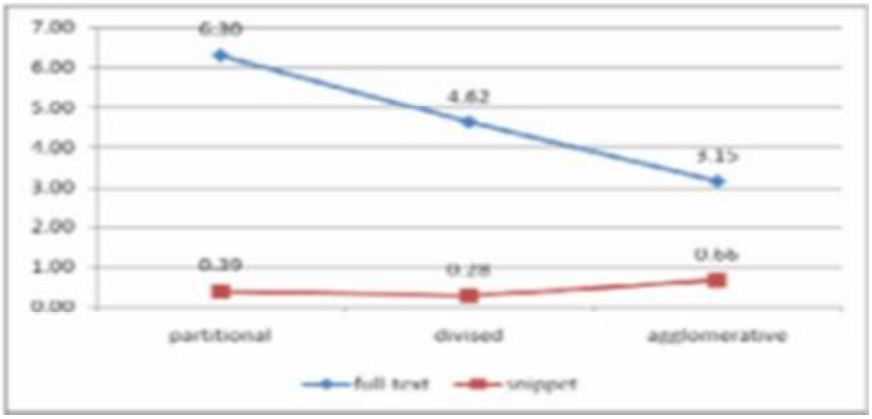
فیلتر سازی: در این مرحله تلفیق اطلاعات مشابه موتورهای جستجوگر و حذف عناوین عینا مشابه و حذف کاراکترهای غیر مجاز، اعمال عملیات منطقی و جبری تحت پوشش (و، یا) صورت می‌گیرد. از طرف دیگر در صورتی که کاربر فیلترینگ بازه زمانی ، سایت‌ها یا گروه‌های پایه را انتخاب کرده باشد، این موارد نیز اعمال می‌گردند.

حذف کلمات عمومی: این قسمت به تشخیص، علامت‌گذاری و حذف کلمات عمومی (بسته به کاربرد) می‌پردازد. در اکثر زبان‌ها لیستی از کلمات عمومی پر تکرار وجود دارد، که در مرحله پیش پردازش آنها را حذف می‌نماییم. عملیات حذف در دو مرحله قبل و بعد از عمل ریشه‌یابی صورت گرفته و نقش بسزایی در کاهش تعداد کلمات قابل پردازش دارد. در مواردی که قصد ایجاد لیبل یا برچسب برای گروه‌ها داریم، جهت بالابردن خوانایی و وضوح برچسب در ابتدا کلمات عمومی را علامت‌گذاری نموده و در مرحله آخر عباراتی که تنها شامل کلمات عمومی هستند یا ابتدا و انتهای آنها با این کلمات است را حذف می‌کنیم. به عنوان مثال کلمه "برای" جزء کلمات عمومی محسوب می‌گردد اما بدون شک خوانایی عبارت "تلاش برای آزادی" از عبارت فاقد کلمه عمومی آن یعنی "تلاش آزادی" بسیار بیشتر است.

تقسیم بندی: هدف از این کار تشخیص محدوده جملات، عبارات و یا لغات قابل پردازش در یک متن است. استفاده از علائم نشانه‌گذاری در زبان HTML کاربرد فراوانی دارد. به عنوان مثال مطالب درون تگ title به راحتی قابل تشخیص و جداسازی هستند.

انتخاب کلمات کلیدی: پس از انجام پیش‌پردازش، فضای محاسباتی و لغات قابل پردازش به میزان چشمگیری کاهش پیدا کرده و تنها تعدادی کلمه کاندید باقی می‌مانند. حال باید به کمک یکی از روش‌های وزندهی به لغات و یا ترکیبی از آنها، ارزش هر کلمه را محاسبه نموده و کلمات نهایی را تشخیص دهیم. مهم‌ترین روش‌های وزن‌دهی در بخش قبلی آورده شده است، حال با توجه به نوع داده‌های موجود باید به انتخاب یکی از متدها پرداخته و با قرار دادن حد آستانه مناسب با داده‌ها، تعدادی از آنها را به عنوان لیبل‌های اصلی انتخاب نماییم.

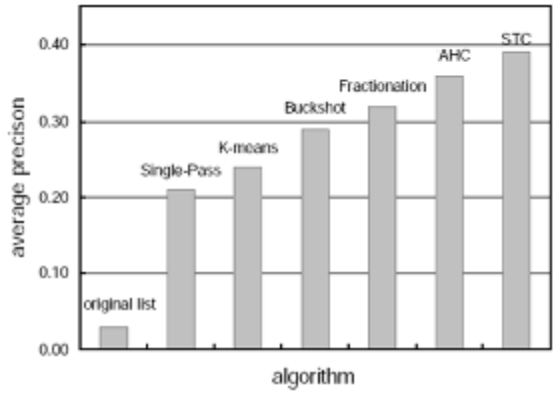
هر چند تمام روشها از مقبولیت و کارایی بالایی در پروژه‌های دسته‌بندی اسناد برخوردارند، اما همچنان ذات داده‌های مورد استفاده اهمیت بالایی دارند. به عنوان مثال در مورد استفاده از قطعات خبری به دلیل کوتاه بودن طول عبارات قابل پردازش فرکانس تکرار پایین است، و یا در مورد اسناد وب، بدون شک اهمیت واژه‌های آورده شده در عنوان و یا با فونت و حالت نوشتاری متفاوت اهمیت بیشتری دارد. عدم توجه به این امر باعث شده است که میزان دقت و کارایی در نمونه‌هایی که کل متن را پردازش می‌کنند نسبت به مواردی که بر روی قطعات کار می‌کنند، در زمانی قابل اغماض بالاتر باشد. این مطلب در شکل (2-4) نشان داده شده است.



شکل(2-4): مقایسه میانگین زمان اجرا و دقت برای تعداد دسته و درخواست مشابه در کل متن و قطعات وب [47]

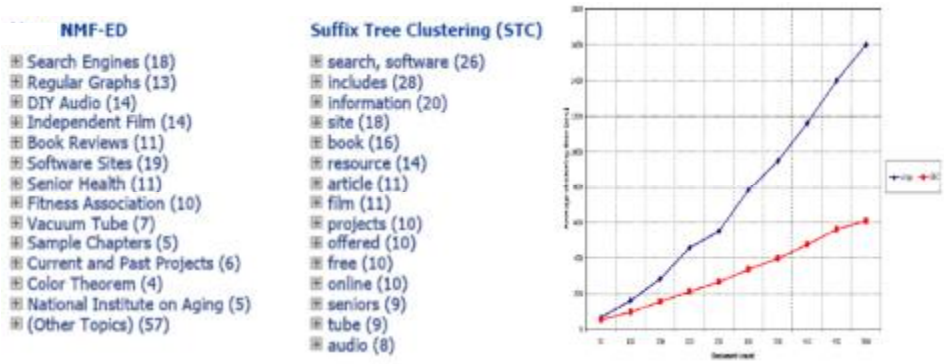
در خوانایی لیبل‌ها، اکثر موارد عبارات با طول بیشتر (بیش از یک کلمه) برای کاربر جذابیت و مفهوم واضح‌تری دارند. در نتیجه اولویت انتخاب با عبارات بوده و باید ضریبی برای افزایش وزن آن‌ها در نظر گرفته شود.

دسته‌بندی داده‌ها: پس از انتخاب کلمات کلیدی نوبت به دسته‌بندی اسناد در گروه‌های مجزا می‌رسد. برای این کار از دو دیدگاه کلی روش‌های مبتنی بر یادگیری ماشین و دسته‌بند می‌توان استفاده کرد. هر چند هر دو مورد به صورت مفصل در بخش قبلی بیان شده‌اند، اما در این بخش مقایسه اجمالی و کلی از موارد دسته‌بندها داشته و نتایج نهایی را نمایش خواهیم‌داد. لازم به ذکر است که در اکثر موارد از ترکیب روشهای دسته‌بندی و یادگیری ماشین استفاده شده و این دو به صورت مکملی برای هم عمل نموده‌اند. نمودار مقایسه‌ای میانگین دقت الگوریتم دسته‌بندی برای 10 مجموعه سند در شکل (2-5) آورده شده است.



شکل(2-5): میانگین دقت الگوریتم‌های دسته بندی برای 10 مجموعه سند [18]

از میان تمام روش‌هایی که در بخش قبل توضیح داده شد، دقت الگوریتم STC در مورد دسته‌بندی داده‌های وب از تمام موارد بالاتر است. استفاده از ساختار درختی، توجه به عبارات و پیچیدگی زمانی کم از مهم‌ترین مزایای این روش به حساب می‌آید.



شکل(2-6): مقایسه روش‌های Lingo , STC راست: پیچیدگی زمانی، چپ: کیفیت و خوانایی لیبل‌ها[12]

شکل(2-6) به صورت خاص دو روش STC و Lingo را در دو معیار پیچیدگی زمانی و کیفیت برچسب‌های تولیدی مقایسه نموده است. همانطور که ملاحظه می‌گردد STC در زمان کوتاه‌تری قادر به دسته‌بندی تعداد داده مشابه است، در حالی که کیفیت، خوانایی و وضوح مطلب در Lingo بسیار بالاتر بوده و از بیان مباحث کلی در آن پرهیز شده است.

در قسمت دسته‌بندی قطعات با دو رویکرد متفاوت مواجه هستیم. در اولین مورد هر سند یا دسته اسناد جدید به تنهایی مورد پردازش قرار گرفته و با انتخاب کلمات کلیدی به دسته‌های جدیدی افراز می‌گردند، در ادامه دسته‌های جدید با موارد قبلی مقایسه شده و موارد مشابه با هم ادغام می‌گردند. روش دیگر انتساب اسناد جدید به گروه‌های قبلی است. که هر کدام از این موارد در جایگاه خود کاربرد دارند.

در انتها نیز عناوین به صورت دسته‌بندی شده به کاربر نمایش داده می‌شوند. برای مرتب‌سازی نمایش می‌توان از الگوریتم‌های دسته‌بندی و یا اولویت‌دهی بر حسب آخرین زمان به روز رسانی یا ایجاد دسته و یا بر حسب تعداد قطعات مرتبط با هر دسته یا برچسب استفاده نمود.

2-4-2 معرفی موتورهای دسته‌بند

در ادامه جدول (2-9)، شامل موتورهای دسته‌بند معروف به همراه الگوریتم‌های به کار رفته در آنها از مرجع [20] آورده شده است (به غیر از موارد مشخص شده). در این بخش تنها به ذکر موتورهای با بیشینه علمی پژوهشی اکتفا کرده و از موارد تجاری صرف نظر کرده‌ایم، چرا که در این موارد هیچ

گونه معیاری برای مقایسه و ارزیابی نهایی وجود ندارد. هرچند به دلیل اهمیت و اعتبار بالایی که موتور vivisimo دارد، در انتها به آن نیز اشاره نموده‌ایم. در ادامه شرح مختصری از نحوه عملکرد چند موتور به عنوان نمونه آورده شده و تصویری از خروجی carrot2 نیز در شکل (2-7) نمایش داده شده است. باید به این موضوع توجه داشت که تمام موتورهای زیر بر اساس درخواست کاربر و ورود کلمه پرس‌وجو عمل می‌کنند.

System name (Alg. alias)	Year	Text features	Cluster labels	Clustering method	On-line demo	Cluster structure
Grouper (STC)	1998	Single words, phrases	phrases	STC	Yes (dead)	Flat
Lassi	2000	Lexical affinities	Lexical affinities	AHC	No	Hieiarchy
CIIRarchies	2001	Single words	Word sets	Graph analysis	Yes (dead)	Hieiarchy
WICE(SHOC)	2002	Single words, phrases	Phrases	SHOC	Yes (dead)	Hieiarchy
Carrot^2 (Lingo)	2003	Frequent phrases	Phrases	Lingo	Yes	Flat
WebCat	2003	Single words	Words	k-Means	Yes (dead)	Flat
AOSearch	2004	Single words	Word sets	AHC	Yes (dead)	Hieiarchy
SnakeT [48]	2004	Approximate sentences	phrases	Approx. sent. coverage	Yes (dead)	Hieiarchy
EigenCluster	2005	Single words	3 salient terms	Divide merge	No	Flat
WhatsOnWeb	2006	Single words	phrases	Edge connectivity	No	Graph
NeSReC[49]	2006	Single words, phrases	phrases	STC	No	Hieiarchy

جدول(2-9): خلاصه‌ای از سیستم‌های پژوهشی دسته‌بندی نتایج جستجو [20]

Grouper

این موتور جستجوگر رابط دسته‌بند اسناد در سرویس جستجوگر چندگانه HuskySearch می‌باشد که در سال 1995 پیاده‌سازی شده است. HuskySearch، نتایج مورد نظر را از چندین موتور جستجو گر معروف دریافت کرده و Grouper، نتایج را به محض دریافت توسط روش دسته‌بند درخت پسوندی (STC) دسته‌بندی می‌نماید. در ادامه رابط کاربری این سیستم به صورت خلاصه شرح داده شده است. کار با ورود درخواست توسط کاربر آغاز می‌گردد. کاربر می‌تواند در این مرحله نحوه برخورد با درخواست خود را مشخص نموده (به عنوان عبارت، کلمات مجزا و ..) و همچنین تعداد و میزان شرکت پذیری هر کدام از موتورهای جستجوگر پایه و تعداد نهایی نتایج جستجو (10-200) را تعیین نماید. به صورت پیش فرض سیستم از 10 موتور استفاده کرده و متعاقبا بین 70 الی 1000 سند را پس از حذف اسناد مشابه دریافت می‌کند. پس از آن نتایج اصلی در قالب تعداد دسته‌های یافته شده و اسناد مرتبط با آنها در صفحه نمایش داده می‌شود. دسته‌ها در یک جدول بزرگ نشان داده شده و هر دسته یک ردیف مجزا را به همراه خلاصه‌ای سند به خود اختصاص می‌دهد. دسته‌ها بر اساس میزان ارتباط تخمین زده شده با درخواست مرتب می‌شوند. خلاصه هر دسته شامل سایز آن (تعداد اسناد) بوده و تلاش می‌کند برای انتقال بهینه محتوای دسته عبارات مشترک میان اسناد و عناوین نمونه را به نمایش گذارد. اعدادی که در پرانتز مقابل هر عبارت آورده شده است نمایانگر درصد اسنادی از دسته است که شامل آن عبارت خاص می‌باشند[50] .

WICE (SHOC)

موتور خوشه‌بند اطلاعات وب[58] از الگوریتمی به نام دسته‌بند آنلاین سلسله مراتبی مفهومی[59] در سیستم خود استفاده نموده است که به صورت محلی داده‌ها را اداره کرده و با موفقیت با سیستم‌های آلفابت بزرگ کنار می‌آید. برای این کار از ساختار آرایه پس‌وندی به جای درخت پسوندی برای استخراج تعداد تکرار عبارات استفاده شده است[51].

[58] Web Iinformation Clustering Engine
[59] Semantic Hierarchical Online Clustering

WebCAT

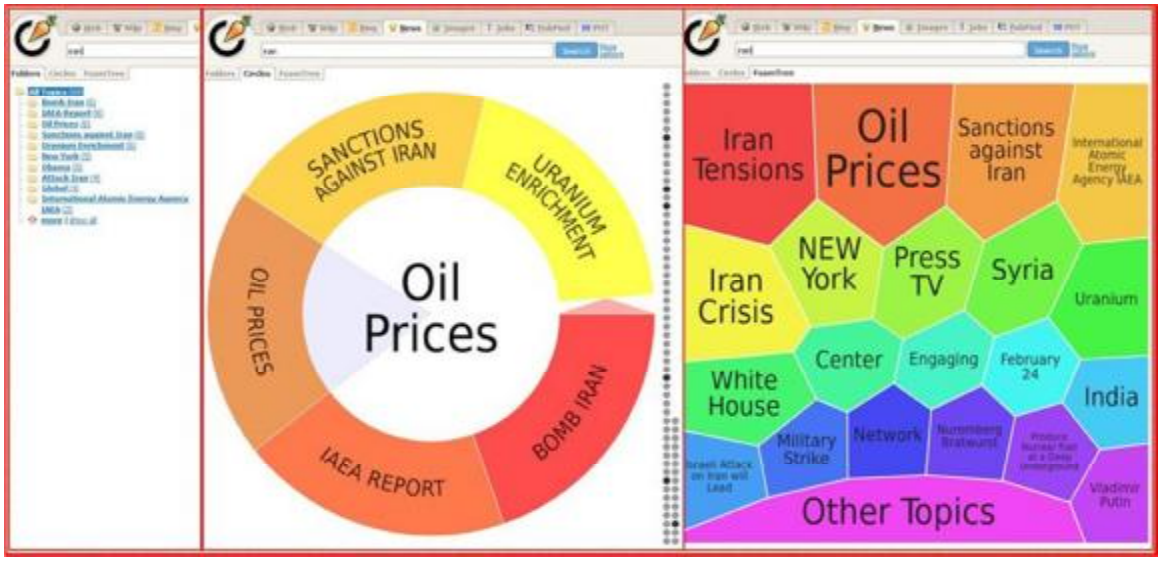
مبنای اصلی این روش که در سال 2003 معرفی شده است، استفاده از نوعی الگوریتم kmeans برای دسته‌بندی داده‌های طبقه‌بندی‌شده می‌باشد. الگوریتم در اصل برای کار با پایگاه‌های داده توسعه داده شده است و در آن از پردازش تعاملی و تعاریف دقیق از تشابه/ عدم تشابه میان اسناد به کمک کلمات طبقه بندی شده استفاده شده است. پیچیدگی زمانی این الگوریتم نسبت به تعداد اسناد، خطی بوده و دارای تعداد تکرار ثابت است[52].

SnakeT

این نام متعلق به سیستم دسته‌بند و الگوریتم مورد استفاده در موتور می‌باشد. ویژگی جالب آن ایجاد سلسله مراتبی از پوشه‌هایی است که احتمالا با یکدیگر همپوشانی دارند. این سیستم ویژگی جدیدی به نام جملات تقریبی شامل عبارات غیر پیوسته (جملات تشکیل شده از عباراتی هستند که در میان آنها فاصله وجود دارد) را معرفی نموده است. پیچیدگی زمانی این الگوریتم برابر با $O(nlog2n+mlog2mp)$ است که در آن n, m, p به ترتیب از چپ به راست نمایانگر تعداد اسناد، تعداد ویژگی‌ها و تعداد لیبل‌ها است[53].

Carrot2

این سیستم در سال 2003 معرفی شده و ترکیبی از چندین الگوریتم دسته‌بندی نتایج جستجو از جمله STC، Lingo ، دسته‌بندی مبتنی بر هوش ازدحامی (بر مبنای نظریه کلونی مورچگان) و تکنیک‌های ساده سلسله مراتبی می‌باشد. اساس کار Lingo استفاده از تکنیک SVD به عنوان مکانیزم اولیه‌ای برای ایجاد خوشه‌ها است. این موتور اغلب تمایل به ایجاد تعداد پوشه بیش از تعداد کل قطعات دارد که نتیجه آن تاثیر منفی بر روی قابلیت و کارایی سیستم است. ازطرف دیگر این سیستم قادر به ادغام لیبل‌های مشابه مانند "دانش و کشف دانش" نبوده و هر کدام را به عنوان یک دسته منفرد در نظر می‌گیرد و این مسئله به صورت سلسله مراتبی اطلاعات اضافی هنگام مرور نتایج به وجود می‌آورد [20].



شکل(2-7): تصویری از خروجی موتور دسته بند carrot2 در مورد درخواست Iran
{ http://search.carrot2.org/stable/search/}

همان‌گونه که در شکل (2-7) مشاهده می‌گردد، نمایش خروجی به سه شیوه متفاوت پوشه‌ای، دایره-ای و ساختار درختی فوم آورده شده است. این کار علاوه بر دید زیبایی‌شناسی، دیدی عینی از نتایج به کاربر خواهد داد. به عنوان مثال در مورد آخر هر چه قسمت رنگی مختص به یک دسته بزرگتر باشد، بی‌شک اهمیت آن نیز بیشتر بوده و وسعت داده بالاتری را پوشش داده است. در نتیجه کاربر به آسانی می‌تواند عناوین مهم و مورد علاقه خود را شناسایی نموده و با مراجعه مستقیم به آنها در زمان خود صرفه‌جویی نماید.

Vivisimo

آخرین موتور مورد بحث، یک مورد تجاری است که هر چند در جدول مقایسه‌ای آورده نشده اما از کارایی بالایی برخوردار می‌باشد. به احتمال زیاد اولین موتور خوشه‌بند تجاری به نام Northern Light متعلق به اواخر دهه 1990 می‌باشد. این موتور شامل تعدادی دسته از پیش تعیین‌شده بوده و نتایج به دست آمده به هرکدام از این دسته‌ها انتساب داده می‌شود. موفقیت بزرگ دیگر در این زمینه موتور vivisimo است که در آن دسته‌ها و لیبل آنها به صورت اتوماتیک استخراج و تولید می‌شوند. این موتور توسط محققان دپارتمان کامپیوتر دانشگاه کارنگی ملون زیر نظر و تحت حمایت‌های مالی

بنیاد ملی علوم در سال 2000 تاسیس شده است. این موتور در سال‌های 2001 الی 2003 جایزه بهترین موتور جستجوگر چندگانه را توسط Search Engine Watch به خود اختصاص داده بود. اساس کار بر مبنای نظریه قدیمی هوش مصنوعی مبنی بر اینکه "یک سیستم دسته‌بند خوب، سیستمی است که دارای توضیحات خوب و خوانا باشد" پایه‌گذاری شده است. نرم افزارهای دسته-بندی اسناد و موتور جستجوگر چندگانه به صورت اتوماتیک نتایج را به محض دریافت (on-the-fly) درون دسته‌های سلسله مراتبی قرار می‌دهند. سرعت این موتور با استفاده از مزایای استانداردهای XML و XSL از ساختاری مدرن بهره می‌گیرد. تا جایی که vivisimo قابلیت رسیدگی به اطلاعات ترابایتی با حداقل هزینه زیر ساختی نسبت به موارد مشابه در بازار تجاری را دارد. از دیگر سیستم‌های تجاری می‌توان به WebClust, KartOOوMooter اشاره نمود[20].

3-4-2 سیستم دسته‌بندی اسناد در مقابل موتورهای دسته‌بند وب

الگوریتم‌های دسته‌بندی برای استفاده در سیستم‌های مدیریت پایگاه داده و سازماندهی صفحات وب در موتورهای جستجو توسعه داده شده‌اند.

با توجه به اینکه اسناد مرتبط با یکدیگر، با درخواست‌های مشابه تطابق دارند، عمل دسته‌بندی اسناد در وب به منزله بهبود اثربخشی و کارایی در رتبه‌بندی اسناد صورت گرفته است. بر این اساس اسناد به گروه‌های مجزا دسته‌بندی شده و گروه‌های با بیشترین میزان تشابه به عنوان خروجی در رتبه‌های بالاتر به صورت سلسله مراتبی به کاربر ارائه داده می‌شود. موتورهای دسته‌بند اسناد به کاربر این اجازه را می‌دهند که تنها دسته‌های مورد نظرش را انتخاب نموده و به مطالعه آن بپردازد. ورودی این سیستم عبارت است از تعدادی نتیجه برگشته از موتورهای جستجو شامل URL، عنوان،تاریخ، سایت منبع و قطعه متنی (متنی کوتاه که به صورت خلاصه محتوای سند را نشان می‌دهد) و خروجی آن نمایش تعدادی دسته لیبل‌گذاری‌شده در ساختاری سلسله‌مراتبی، درخت‌مانند یا پارتیشن‌های مسطح است. در جدول (2-10) عمده تفاوت‌های میان این سیستم‌ها آورد شده است.

Clustering type	Cluster label	Cluster computation	Input data	Cluster intersection	GUI
Search results clustering	Pre Defined	Online	Snippets	Overlapping	Yes
Document clustering	Centroid	Offline	Documents	Disjoint	No

جدول(2-10): دسته‌بندی نتایج جستجو در مقابل دسته‌بندی‌های عمومی و سنتی متون [54]

جدول فوق تفاوت‌های کلی میان دسته بندهای عمومی و سنتی متون و اسناد وب را نشان می‌دهد. منظور از دسته‌بند عمومی به عنوان مثال سیستم دسته‌بندی عناوین کتب درون یک کتابخانه الکترونیک می باشد. برچسب اصلی هر دسته که به گونه‌ای نماینده اعضا مجموعه محسوب می‌گردد در اسناد وب، نوعی برچسب مفهومی شامل یک عبارت، لغت و یا توصیفی به زبان طبیعی گفتاری است در حالی که در سیستم‌های دسته‌بندی متون اغلب از یک ویژگی آماری مانند میانگین نمایش عددی مفاهیم استفاده می‌گردد در نتیجه خوانایی و قابلیت درک برچسب در سیستم‌های اول اهمیت بالاتری دارند. به همین دلیل در صورت اضافه‌شدن یک سند جدید به اسناد در وب، برچسب همچنان بدون تغییر باقی می‌ماند اما در سیستم‌های متنی، میانگین داده‌های دسته هرچند بسیار کم و نامحدودتغییر می‌یابد. در سیستم‌های عمومی داده‌ها از پیش در یک پایگاه ذخیره شده و برای پردازش آماده گشته‌اند در حالی که اسناد وب در هر لحظه بنا به درخواست کاربر و یا گذر زمان به صورت مداوم و پیوسته تغییر کرده و باید عملیات به صورت آنلاین بر روی آنها صورت گیرد. نوع داده-های مورد پردازش نیز در وب، قطعات بدست آمده از موتورهای جستجوگر است که اغلب شامل عنوان، خلاصه‌ای کوتاه، تعداد تگ معنایی، آدرس و وب‌سایت مرجع است. در حالی‌که در دسته‌بندی متون از کل متن و محتوای موجود که اغلب طولانی است، استفاده می‌کنیم. تعداد دسته‌های نهایی در متون مشخص بوده و تنها باید انتساب اسناد به گروه ها صورت گیرد ولی در وب این تعداد نامشخص بوده و هر داده می‌تواند در زمان واحد به بیش از یک گروه با درصدهای متفاوت تعلق داشته باشد. در نهایت برای سیستم‌های وب نیاز به یک واسط کاربری برای دریافت اطلاعات و نمایش خروجی و به

صورت کلی تعامل با کاربر وجود دارد ولی در دسته‌بندی متون انتساب داده‌ای به منزله سهولت دسترسی خود سیستم بوده و نهایتا نیاز به یک صفحه برای نمایش اطلاعات داریم.

4-4-2 ارزیابی عملکرد موتورهای دسته‌بندی

6 معیار کلی برای ارزیابی عملکرد سیستم‌های سیستم های بازیابی اطلاعات وجود دارد که عبارتند از: پوشانندگی درخواست[60]، فراخوانی[61] ، دقت[62]، زمان پاسخ[63]، تلاش کاربر[64] و فرمت خروجی[65]. از آنجا که این معیارها در سال 1973 [55] معرفی گشته‌اند بسیاری از آنها ارزش خود را تا حدی کاهش داده‌اند، به عنوان مثال میزان تلاش کاربر برای دسترسی به نتایج و ارسال درخواست خود با توجه به طراحی رابط گرافیکی بسیار کاهش یافته است، اما همچنان برای سیستم‌های کنونی مورد استفاده قرار می‌گیرد. اما در عین حال توجه به جنبه‌های زیر نیز بسیار اهمیت دارد[54].

ترکیب شاخص‌های وب: زمانی که یک درخواست ایجاد می‌شود، شاخص نتایج به کمک ربات‌ها و عنکبوت‌ها ایجاد می‌گردد نه توسط خود صفحات. در نتیجه ترکیب شاخص‌های وب تاثیر مهمی بر عملکرد موتور جستجو دارد.

سه جزء کلی وجود دارد که در ساختار آرایش شاخص وب اهمیت دارد که عبارتند از پوشش وب، زمان به‌روزرسانی و بخش‌های شاخص‌گذاری (به عنوان مثال عناوین به انضمام چند خط اول، یا کل صفحه وب). از سوی دیگر محدوده پوشش بزرگتر، به روز رسانی مکرر و نمایه‌سازی کل متن لزوما به معنی بهتر بودن موتور جستجوگر نیست.

قابلیت‌های جستجو: موتور جستجوگر باید امکانات اساسی که کاربر با آنها آشنایی دارد را داشته باشد که شامل عبارات منطقی بولین، جستجو عبارت، اختصارات و تسهیلات محدود ساز (محدودیت تعداد و حجم اطلاعات) باشد.

---

[60] Coverage
[61] Recall
[62] Precision
[63] Response time
[64] User effort
[65] Form of output

عملکرد میزان بازیابی: این عملکرد به صورت سنتی شامل سه پارامتر دقت، فراخوانی و زمان پاسخ می‌باشد.

مجموعه اسناد مرتبط با یک پرس وجو را با {Relevant} نشان می‌دهیم، و مجموعه اسناد بازیابی شده را با {Retrieved} نشان می‌دهیم. مجموعه اسنادی که هم مرتبط هستند و هم بازیابی شده-اند با {Relevant} ∩ {Retrieved} نشان داده می‌شوند.

دو مقیاس اصلی برای تشخیص کیفیت بازیابی متن وجود دارد:

- دقت: در صد اسناد بازیابی شده که در واقع به پرس‌وجو مرتبط هستند (برای مثال پاسخ‌های «درست»)

$$precision = \frac{|\{\text{Re}levant\} \cap \{\text{Re}trieved\}|}{|\{\text{Re}trieved\}|} \qquad (4\text{-}2)$$

- فراخوانی : در صد اسنادی که به پرس‌وجو مرتبط هستند و در واقع بازیابی شده‌اند.

$$recall = \frac{|\{\text{Re}levant\} \cup \{\text{Re}trieved\}|}{|\{\text{Re}trieved\}|} \qquad (5\text{-}2)$$

- معیار F: که ترکیبی از دو روش فوق بوده و به صورت زیر محاسبه می‌گردد.

$$F - measure = \frac{2 * Recall * Precision}{Recall + Precision} \qquad (6-2)$$

فرمت و شمای خروجی: این جزء ارزیابی باید از دو دیدگاه مورد بررسی قرار گیرد. یکی از موارد تعداد گزینه‌های خروجی و دیگری توجه به محتوای حقیقی خروجی است.

تلاش کاربر: این گزینه بیشترین ارتباط را با نحوه آماده سازی مستندات و رابط کاربری تهیه شده دارد. از آنجا که تعداد بسیار زیادی موتور جستجوگر وجود دارد، باید به میزان سهولت کار برای کاربر به منزله جذاب نمودن هرچه بیشتر آن توجه نمود.

5-2 دسته‌بندی اسناد خبری در وب

با توجه به ماهیت پویای اطلاعات در صنعت خبرگزاری و به‌روزرسانی ثانیه‌ای اخبار در بستر وب در ساعات مختلف شبانه روز و همچنین عدم تطابق زمانی در مناطق مختلف جغرافیایی، نیاز به مراجعه مکرر به وب‌سایت‌ها برای مطلع‌شدن از جدیدترین اخبار وجود دارد. از سمت دیگر به دلیل عدم امکان جمع‌آوری اخبار در ایام تعطیلات و ساعات غیر اداری و وجود تعداد زیاد سازمان‌های خبری معتبر در جهان، طراحی سیستمی اتوماتیک برای این منظور لازم به نظر می‌رسد.

در برخی موارد شاهد انتشار اخبار ضد و نقیض بسیاری هستیم و متاسفانه به دلیل گستردگی شبکه نت و دسترسی تمام افراد به آن، امکان نظارت مراجع اجرایی بر روی تمام موارد وجود ندارد، لذا باید سیستمی جهت تطبیق و ارزش‌دهی به اخبار تعبیه گردد. از طرف دیگر در بسیاری موارد مشاهده می‌گردد که اخبار منتشر شده مکمل اخبار قدیمی بوده و باید به منظور شفاف‌سازی بیشتر خبر از ترکیب، ادغام و نهایتا جمع‌بندی آنها استفاده نمود.

در نتیجه می‌توان این چنین بیان نمود که تضاد عمیقی میان حجم اخبار و رشد سریع منابع قابل جستجو و عدم توانایی در کنترل و زمانبندی ورود این اطلاعات به شبکه و توان، زمان و ساعات کاری نیروی انسانی وجود دارد و برای مقابله با آن نیاز به طراحی و پیاده‌سازی سیستمی برای دسته‌بندی اخبار انتشار یافته در وب وجود دارد. در این بخش به صورت خاص اسناد خبری را تحلیل و آنالیز کرده و قصد طراحی سیستمی به منزله دسته‌بندی قطعات خبری به دست آمده از موتورهای جستجو چندگانه داریم. بیشترین کاربرد این طرح، استفاده از آن در خبرگزاری‌ها و سازمان‌های دولتی به منزله ایجاد مزیت رقابتی در میان دیگر سازمان‌ها در زمان و اعتبار دسترسی و بازپخش اخبار و دسته‌بندی بدون ناظر آنها بسته به نمودار خبری مختص سازمان است، هر چند در ادامه می‌توان از این روش به منزله شخصی‌سازی اسناد خبری برای کاربران این صفحات نیز استفاده نمود.

1-5-2 تفاوت اسناد وب و اسناد خبری

دسته‌بندی اسناد و نتایج حاصل از جستجو در وب یکی از شاخه‌های دسته‌بندی اسناد متنی است با این تفاوت که در وب با حجم بالا و غیر قابل کنترل، نظارت و پیش‌بینی صفحات مواجهیم. علاوه بر اینکه احتمال تکراری بودن محتوای صفحات یا تکرار محتوای یک سند در وب وجود دارد، از سمت دیگر با تغییر مداوم مفهوم و متن درون صفحات وب مواجه هستیم. داده های موجود در صفحات وب اغلب به صورت بدون ساختار بوده و از ناهمگونی بالایی برخوردار بوده و در اکثر موارد شامل اطلاعات اضافی بسیاری مانند اتصالات درون و برون صفحه‌ای، ابر اطلاعات و ... می‌باشند.

بنا به دلایل فوق در سیستم‌های دسته‌بندی وب به سیستم‌هایی با سرعت پردازش بالا، قابلیت انعطاف‌پذیری در مقابل تغییرات مداوم محتوا و کاربرگرا نیازمندیم. تعداد دسته اولیه به ازاء هر خبرگزاری ثابت است، حال آنکه هیچ استاندارد کلی که تمام موسسات از آن پیروی کنند وجود ندارد. از طرف دیگر هر خبر می‌تواند در چندین حوزه تاثیرگذار باشد. بدلیل اهمیت اعتبار و صحت اخبار انتشاری، تنها باید به سایت‌های معتبر مراجعه نمود، هر چند تعیین و ردگیری سایت‌های پر بازدید و معتبر ثانویه و افزودن آنها به سیستم نیز امکان‌پذیر است.

هر عنوان خبری مربوط به یک شخصیت، مکان و زمان به خصوص است که این موضوع اهمیت تشخیص موجودیت‌های اسمی[66] را بیشتر می‌کند، از طرف دیگر هر حادثه بلافاصله پس از به وقوع پیوستن توسط اکثر خبرگزاری‌ها انتشار می‌یابد (حجم اخبار مشابه در بازه زمانی کوتاه) و در اکثر موارد از نوع نوشتار رسمی (زبان و قواعد) استفاده می‌شود.

[٦٦] Name Entities

2-5-2 چالش‌های موجود در وب

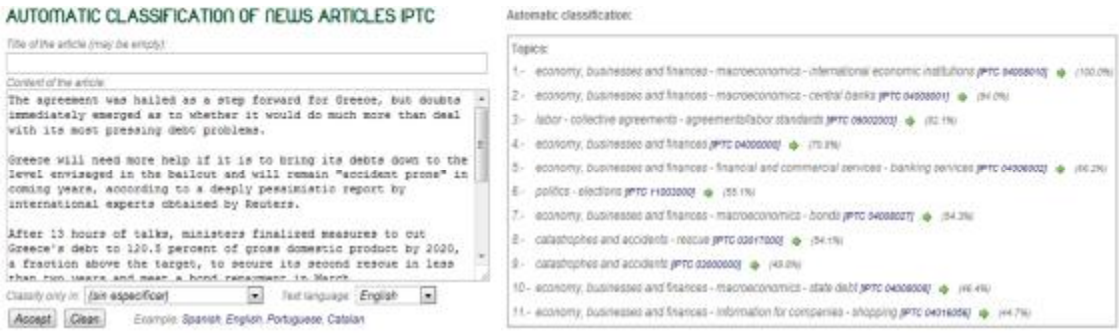
برای انتشار اخبار صحیح باید اعتبار سنجی (به کمک سایت‌های اولیه معتبر) انجام گیرد. وجود اختلاف زمانی بدلیل موقعیت‌های جغرافیای متفاوت و عدم امکان پوشش 24 ساعته اخبار بدلیل ساعت کاری و نیرو و توان محدود انسانی، امکان دسترسی دستی به اخبار دقیق در کوتاه‌ترین زمان ممکن را دشوار می‌سازد که این امر می‌تواند در اخذ تصمیمات، مدیران را با چالش مواجه سازد. برخی از مهم‌ترین چالش‌های موجود در این زمینه در زیر لیست شده‌اند.

- تطبیق اخبار و اعتبار سنجی
- توان، زمان و ساعات کار محدود نیروی انسانی
- اطلاع دقیق و سریع از اخبار جهت اخذ تصمیم
- عدم وجود استاندارد خبری و وجود شاخه‌های غیر مشابه
- پراکندگی زمانی بدلیل تفاوت در مختصات جغرافیایی و میزان تاثیر پذیری

2-5-3 بررسی دیدگاه‌های متفاوت

ایجاد فضایی برای تشخیص گروه خبر

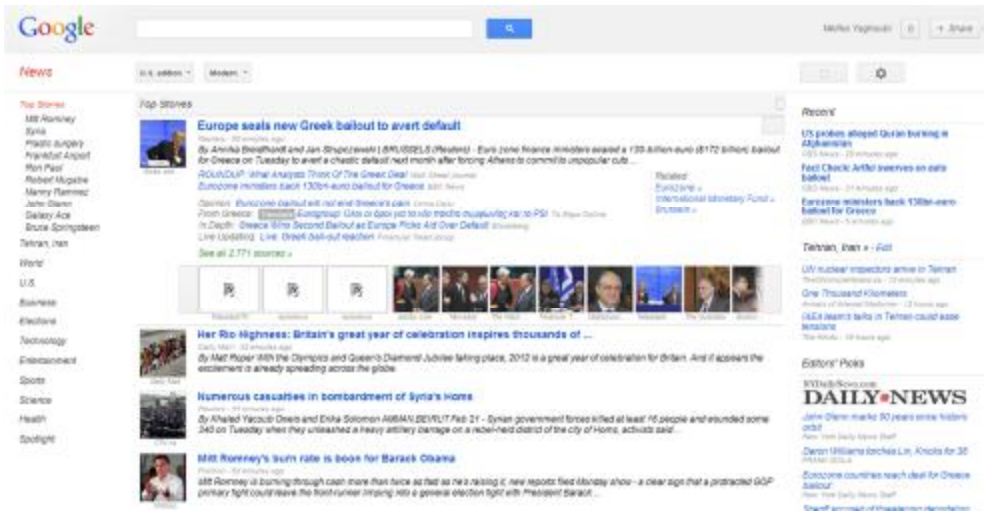
اولین دیدگاه در دسته‌بندی اخبار تعریف یک چارت سازمانی دقیق به همراه موجودیت‌های ارزشمند در هر کدام از دسته‌ها است. بدین صورت می‌توان با وارد نمودن متن و عنوان خبر میزان تعلق آن به گروه‌های مختلف استاندارد و از پیش تعین شده را تشخیص داد. نتیجه این رویکرد در شکل (2-8) آورده شده است.



شکل(2-8): ایجاد فضایی برای تشخیص گروه خبر توسط موتور دسته بند نتایج
چپ: عنوان یا متن خبری ورودی، راست: دسته های متعلق به خبر و درصد تعلق
{http://showroom.daedalus.es/}

دسته‌بندی موضوعات خبری

یکی دیگر از دیدگاه‌های مطرح دسته‌بندی موضوعات خبری به عناوین داغ و جدید روز، عناوین پر بازدید، عناوین اخیر و ... است که در این موارد بیشترین پردازش و آنالیز بر روی اطلاعات اضافی خبر (زمان انتشار، تعداد کلیک، آنالیز لینک های درونی و بیرونی و ...) بوده و کمتر به مفاهیم و محتویات خبر پرداخته می‌شود. در این زمینه سایت‌های زیادی مانند google news و yahoo news کار کرده‌اند. اما این شیوه نیز نمی‌تواند دیدی دسته‌بندی شده و آسان در اختیار کاربر قرار دهد. این نمونه در مرجع [54] نیز بیان شده و در شکل (2-9) نمایش داده شده است.

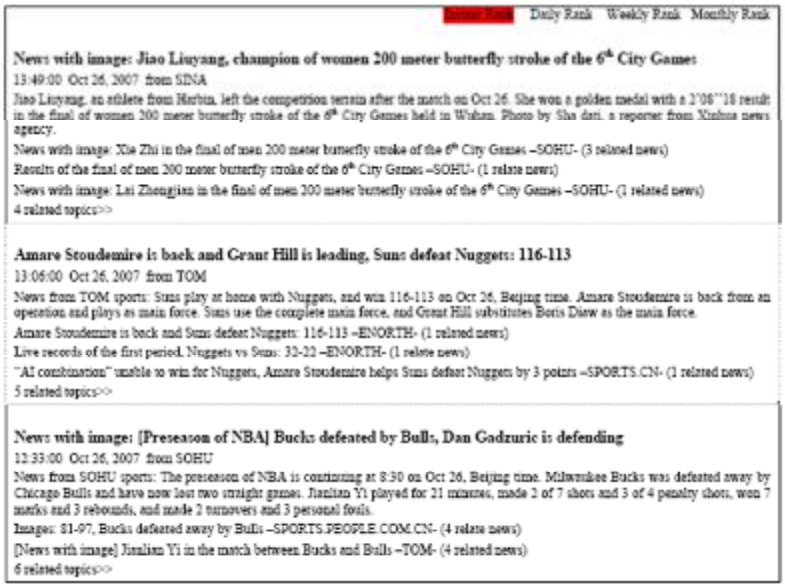


شکل(2-9): دسته بندی موضوعات خبری توسط موتور دسته بند نتایج
{http://news.google.com/}

ایجاد و ردیابی جریان خبری

در اغلب موارد مردم علاوه بر جستجو کلمه‌ای عبارات تمایل دارند از امور جاری سر در بیاورند. یکپارچه سازی اطلاعات جامع در مورد موضوعات خبری امری لازم است که آنرا "جریانات خبری" می‌نامیم و شامل پس‌زمینه، تاریخچه، پیشبرد جریانات، نقدها و نظرات مختلف پیرامون آن است. به طور سنتی این کار توسط ویرایشگران خبری در خبرگزاری‌ها و مراجعه به بایگانی و آرشیو خبری صورت می‌گیرد که بسیار زمانبر بوده و امکان بروز بودن و دسترسی آنی به آنها وجود ندارد. نمونه یک جریان خبری در شکل (2-10) آورده شده است.

این کار به ایجاد یک تاریخچه غیر مدون از اخبارها در بازه زمانی طولانی می‌پردازد. به عنوان مثال تمامی جزئیات و اخبار مربوط به مسابقات المپیک لندن از روز نخست تا امروز. این سیستم‌ها هنوز به صورت عمومی قابل استفاده نبوده و در حال حاضر به صورت پژوهش‌های علمی [58 ,57] به آن پرداخته می‌شود.


Daily Rank Weekly Rank Monthly Rank

News with image: Jiao Liuyang, champion of women 200 meter butterfly stroke of the 6th City Games

13:49:00 Oct 26, 2007 from SINA

Jiao Liuyang, an athlete from Harbin, left the competition terrain after the match on Oct 26. She won a golden medal with a 2'06''16 result in the final of women 200 meter butterfly stroke of the 6th City Games held in Wuhan. Photo by Sha dan, a reporter from Xinhua news agency.

News with image: Xie Zhi in the final of men 200 meter butterfly stroke of the 6th City Games –SOHU- (3 related news)

Results of the final of men 200 meter butterfly stroke of the 6th City Games –SOHU- (1 relate news)

News with image: Lai Zhengjian in the final of men 200 meter butterfly stroke of the 6th City Games –SOHU- (1 related news)

4 related topics>>

Amare Stoudemire is back and Grant Hill is leading, Suns defeat Nuggets: 116-113

13:06:00 Oct 26, 2007 from TOM

News from TOM sports: Suns play at home with Nuggets, and win 116-113 on Oct 26, Beijing time. Amare Stoudemire is back from an operation and plays as main force. Suns use the complete main force, and Grant Hill substitutes Boris Diew as the main force.

Amare Stoudemire is back and Suns defeat Nuggets: 116-113 –ENORTH- (1 related news)

Live records of the first period, Nuggets vs Suns: 32-22 –ENORTH- (1 relate news)

"All combination" unable to win for Nuggets, Amare Stoudemire helps Suns defeat Nuggets by 3 points –SPORTS.CN- (1 related news)

5 related topics>>

News with image: [Preseason of NBA] Bucks defeated by Bulls, Dan Gadzuric is defending

12:33:00 Oct 26, 2007 from SOHU

News from SOHU sports: The preseason of NBA is continuing at 8:30 on Oct 26, Beijing time. Milwaukee Bucks was defeated away by Chicago Bulls and have now lost two straight games. Jianlian Yi played for 21 minutes, made 2 of 7 shots and 3 of 4 penalty shots, won 7 marks and 3 rebounds, and made 2 turnovers and 3 personal fouls.

Images: 81-97, Bucks defeated away by Bulls –SPORTS.PEOPLE.COM.CN- (4 relate news)

[News with image] Jianlian Yi in the match between Bucks and Bulls –TOM- (4 related news)

6 related topics>>

شکل(2-10) ایجاد جریان خبری توسط موتور دسته بند نتایج [58]

فصل سوم: مروری بر تحقیقات

1-3 مقدمه

در این فصل به معرفی مجموعه‌ای از پژوهش‌های انجام شده در حوزه‌های مورد بحث می‌پردازیم. بدین منظور، شماری از مطالعات صورت گرفته در زمینه پردازش و دسته‌بندی متون، موتورهای جستجوگر و دیگر فیلدهای مرتبط با پروژه را معرفی خواهیم نمود. لازم به ذکر است که در این بخش از بیان جزئیات و دوباره‌گویی عملکرد روش‌ها و الگوریتم‌هایی که در فصل قبل بیان گشته‌اند، چشم-پوشی کرده و تنها به بیان کلیات خواهیم پرداخت.

نحوه استفاده از تکنیک خوشه بندی با نظارت برای دسته‌بندی نامه‌های دریافتی در سامانه 137 شهرداری تهران در [26] شرح داده شده است، که در آن به کمک الگوریتم بیزین عمل دسته‌بندی صورت گرفته است. در این پروژه بعد از پیش‌پردازش داده‌های خام به منظور تبدیل آنها به فرمت مناسب برای آموزش و وزندهی به عبارات، پیام‌های فارسی ارسالی از طرف شهروندان از طریق موبایل به گروه‌های از پیش تعیین شده نسبت داده شده‌اند.

در مقاله [30] از یک روش آماری ترکیبی، برای استخراج کلمات کلیدی اسناد فارسی استفاده شده است. در این روش پس از اعمال پیش‌پردازش‌های مورد نیاز ویژگیهای آماری برای کلمات مختلف محاسبه شده و با استفاده از فازی‌سازی و اعمال قواعد فازی، کلمات کلیدی محتمل، انتخاب می-شوند.گام بعدی محاسبه رخداد همزمان 4 پیشین و پسین کلمات کلیدی محتمل، با کلمات تکرار شونده، 5 در جملات سند است. در ادامه با اعمال حد آستانه مشخص وفقی روی رخداد همزمان کلمات، عبارات دو کلمه‌ای تشخیص داده می‌شوند.

در این پروژه بر روی سه گروه کلمه عمومی (کلماتی که در کلیه اسناد کلیه کلا سها پخش شده‌اند)، کلمه خاص (کلماتی با فراوانی زیاد فقط در کلیه اسناد یک کلاس و فراوانی کم در سایر کلا سها) و کلمات کلیدی (عبارت یا کلمه مهم و دارای فراوانی زیاد در سند خاص) مانور زیادی داده شده است. به همین منوال از روش‌های وزن‌دهی متنوعی نیز در سه گروه کلمه فوق استفاده شده است.

مقاله [44] با هدف یافتن و استخراج کلمات کلیدی مستندات فارسی، به منظور طبقه‌بندی کارآمد آنها در موتورهای جستجو به ترکیب چند تکنیک ریشه‌یابی و خلاصه سازی اسناد پرداخته است. پس از ورود هر متن عملیات حذف وندها و کلمات اضافه و افعال صورت گرفته و به هر جمله امتیاز مشخصی داده می‌شود. سپس به حذف کلمات هم‌معنا و تعمیم کلمات به کمک دیکشنری‌های جانبی پرداخته می‌شود. در تعریف فاکتور اهمیت برای کلمات با در نظر گرفتن محدوده‌ای مشخص میزان تکرار بین حد فرکانس پایین و بالا بر اساس تعداد کلمات موجود در متن را محاسبه نموده و منبع کلمات مهم را ایجاد می‌کنیم. در نهایت درجه اهمیت هر کلمه توسط حاصل ضرب فرکانس تکرار کلمه در سند در تعداد جملات شامل کلمه در سند محاسبه می‌شود. در مرحله بعد به منزله یافتن فاکتور اهمیت جمله، با تعیین حدی به عنوان حداکثر فاصله دو کلمه مهم، یک بخش در جمله را پیدا می‌کند که به وسیله کلمات مهمی که بیش از Lکلمه غیرمهم بینشان قرار ندارد، جدا شده‌اند. سپس تعداد کلمات مهم موجود در این بخش را شمرده و جذر این تعداد را بر تعداد کل کلمات موجود درآن بخش تقسیم نموده و فاکتور اهمیت جمله را بدست می‌آورد. بعد از آن به حذف کلمات هم معنی با استفاده از لغت‌نامه و انتخاب کلمه جایگزین پرداخته می‌شود که نتیجه آن همگونی و ساده سازی متون می‌باشد. در این پروژه از تعداد 100 متن غیر استاندارد در 5 گروه و مقایسه آن با روش سنتی و دستی استفاده شده است.

مقاله [49] با گرفتن در خواست از کاربر، جمع‌آوری تکه‌های خبری توسطAltavista به مدیریت نتایج جستجو به منزله جلوگیری از مشاهده نتایج تکراری توسط کاربر پرداخته و در نهایت بر مبنای لیبل‌گذاری، عمل دسته بندی را انجام می‌دهد. در این روش پس از دریافت نتایج اولیه از موتور جستجوگر به استخراج عناوین کاندید و اولویت‌بندی عنوان‌ها پرداخته و در نهایت حاصل دسته‌بندی نمایش داده می‌شود. تمام کلمات غیر مهم بدون نیاز به ریشه‌یابی و همچنین عناوین با کمتر از طول 3 کلمه حذف گشته و با درجه‌بندی عناوین کاندید از ریزاسناد خبری، عنوان برتر انتخاب می‌گردد. در

نتیجه عمل دسته‌بندی به جای استفاده از تشابه میان متون، از تشابه لیبل‌های انتخابی استفاده می-کند.

در مرجع [56] رویکردی جدید در یافتن و دسته‌بندی موارد مشابه به کمک روش SVM در اخبار فارسی مورد بررسی قرار گرفته است. از آنجا که مطالعه اخبار مشابه در وب‌سایت‌ها و اعتبار سنجی آنها برای کاربران بسیار زمانبر است، سیستم دسته‌بندی برای این منابع در نظر گرفته شده و از عنوان و توضیحات اخبار برای دسته بندی استفاده شده است. در نتیجه در پایگاه داده مربوطه آدرس، تاریخ و کلمات کلیدی ذخیره می‌شود. مرحله اول پروژه به استخراج ویژگی متن یک خبر و ذخیره‌سازی آن در پایگاه داده محلی پرداخته و مرحله دوم به استخراج اخبار مشابه بر اساس ویژگی‌های مشخص شده در دیگر صفحات می‌پردازد. این پروژه بر مبنای درخواست کاربر بوده و عمل پالایش و حذف کلمات عمومی در آن به دو صورت آفلاین هنگام پویش اخبار از صفحات وب و آنلاین به معنای حذف کلمات اضافی در هنگام دریافت درخواست کاربر صورت می‌گیرد. در مورد پیدا کردن قطعات مشابه اخبار نیز اولویت بندی قطعات اخبار به صورت تشابه کامل با تمام کلمات درخواست، تشابه با تعداد کلمات منهای یک و ... انجام می‌گیرد.

در اغلب موارد مردم علاوه بر جستجو کلمه‌ای، تمایل دارند از امور جاری با خبر شوند. یکپارچه سازی اطلاعات جامع در مورد موضوعات خبری امری است که آنرا "جریانات خبری" می‌نامیم و شامل پس زمینه، تاریخچه، پیشبرد جریانات، نقدها و نظرات مختلف پیرامون موضوع بوده و مقاله [58] به آن پرداخته است. به طور سنتی این کار توسط ویرایشگران خبری در خبرگزاری‌ها و مراجعه به بایگانی صورت می‌گیرد. در این مقاله الگوریتم سه مرحله‌ای تشخیص موضوع، ردیابی خبر و ادغام موضوعات با یکدیگر، برای ایجاد روند جریانات اخبار در وب پیشنهاد شده است. این سیستم چندین جریان خبری شامل " خبرهای تازه[67]"، "خبرهای داغ[68]" و "چه خبر[69]" را پوشش می‌دهد که پرسیدن آنها

[٦٧] What’s new
[٦٨] What’s hot
[٦٩] What’s going on

به صورت درخواست نتیجه‌ای در موتورهای جستجو نخواهد داشت. عناوین خبری اصولا برای چند روز اهمیت داشته و سپس در مدتی کوتاه ارزش خود را از دست می‌دهند، در حالی که جریانات خبری در مدت زمان طولانی‌تری پایداری داشته و به رکوردهای قابل توجهی تبدیل می‌شوند.

مقاله [59] به بررسی روش‌های مختلف یادگیری ماشین برای دسته‌بندی اتوماتیک اسناد به زبان فارسی پرداخته و دو روش SVM و KNN را با یکدیگر مقایسه نموده است. در این مقاله نیز پس از انجام پیش پردازش و ایندکس گذاری داده‌ها به روش TFIDF و TFDRF، کلمات مهم انتخاب شده‌اند. از 800 سند موجود در همشهری متعلق به 8 دسته مجزا جهت آموزش و تست سیستم استفاده شده است. بهترین نتیجه در روش SVM با کرنل polynomial و در روش KNN مجاورت یک همسایه با معیار شباهت Cosine تشخیص داده شده است. این مفاله بیشتر بر مبنای مقایسه روش‌ها و توجه به ساختار کلمات در زبان فارسی عمل نموده است.

در مقاله [60] کار استخراج کلمات کلیدی بر روی جملات و عبارات غیر رسمی موجود در شبکه‌های اجتماعی به منزله مدیریت تبلیغات و علاقه‌مندی‌های کاربران است. این ریز اسناد اغلب شامل مجموعه کلمات بیانگر وضعیت، علایق و اخبار شخصی برای ارائه به دوستان است. به دلیل طول کوتاه اغلب TF برابر با 1 است، در نتیجه این تعداد تکرار کم، فرکانس کارایی مناسبی ندارد. در این روش تمرکز بر روی موقعیت نسبی کلمه در عبارت، طول عبارت، تعداد کل کلمات و ارزش کلمه، بر حسب نوع نوشتار کلمه با حروف بزرگ یا کوچک است. در نهایت کلمات با امتیاز بالاتر از حد آستانه مشخص، انتخاب می‌شوند. در این پروژه از 1830 عبارت موجود در شبکه اجتماعی "فیس بوک" استفاده شده است.

در مرجع [61] یک دسته‌بند اسناد فارسی با استفاده از الگوریتم K-NN پیشنهاد شده و دلیل استفاده از آن نیز موفقیت زیاد در دسته‌بندی اسناد مشابه در زبان انگلیسی عنوان گشته است. در الگوریتم پیشنهادی، از 540 متن فارسی برگرفته از روزنامه‌های آنلاین ایران و جام‌جم، به عنوان داده آموزشی و از 60 متن مشابه به عنوان داده آزمایشی استفاده شده است. متنها در 6 دسته از پیش

تعریف شده قرار گرفته و اندازه‌گیری مشابهت اسناد که به وسیله برداری از فرکانس خصیصه‌ها در هر سند نمایش داده می‌شوند، از طریق محاسبه شباهت کسینوسی صورت پذیرفته است. برای ارزیابی کارایی، میزان دقت و یادآوری مورد بررسی قرار گرفته و نتایج به دست آمده حاکی از آن است که K-NN قادر است با تعداد کم داده آموزشی به خوبی کار نماید.

یک روش سریع برای خوشه بندی نتایج جستجوی وب نیز در مقاله [62] معرفی شده که در آن از خوشه‌بندی به عنوان روشی هوشمند برای بهبود نمایش نتایج جستجو برگشتی از موتورهای جستجوی وب استفاده شده است. در ابتدا به کمک روش برداری، ماتریس عبارت-متن ایجاد شده و در نهایت با تکیه بر مفهوم دریافتی از موضوع خوشه بندی صورت می‌گیرد. این پروژه نیز مبتنی بر درخواست کاربر بوده و در آن اطلاعات مورد نیاز از موتورهای جستجوگر بدست می‌آیند که هر نتیجه جستجو شامل عنوان، توضیحات، خلاصه و آدرس با ضرایب اهمیت متفاوت است. وزن‌دهی نهایی لغات نیز بر اساس محل قرار گرفتن لغت و تعداد دفعات تکرار آنها صورت می‌گیرد. در این پروژه به کلمه ای لغت کاندید گفته می‌شود که در بخش عنوان یا توضیحات بوده و در بیشتر از دو بخش نتایج دیده شود. در ادامه مباحث بدست آمده از جستجو خوشه بندی شده به ترتیب تعداد سند موجود در هر کدام که نشانگر اهمیت دسته است، نمایش داده می‌شوند. عدم وابستگی به مقادیر آستانه جز مزایای این روش محسوب می‌گردد.

[63] به دسته‌بندی نتایج جستجو بر مبنای ویژگی‌های مدارک و امکان‌سنجی استفاده از الگوریتم-های خوشه‌بندی مختلف در سطح وب پرداخته است. از آنجا که بازآرایی و سازماندهی مدارک همواره بر اساس ویژگی‌های هر مدرک صورت می‌پذیرد ، بازسازی نتایج جستجو نیز منوط به ویژگی‌های مذکور است. ویژگی‌های مدارک در این پروژه به دو گروه ردیف اول و ردیف دوم دسته‌بندی شده است. در این مقاله، با اشاره به انواع ویژگی‌های مطرح برای مدارک و اشاره به این نکته که خوشه‌بندی یکی از روشهای رایج در دسته‌بندی نتایج جستجوست، تلاش شده تا فصل مشترک میان این ویژگی-ها و نیز روشهای مختلف خوشه‌بندی مشخص گردد. این در حالی است در سطح وب، امکان محدود

ساختن نتایج جستجو به تاریخ، عنوان، دامنه، قالب مدرک، در نظر گرفتن پیوندهای فرامتن در نتایج جستجو و نیز حتی جستجو در پی ناشر اثر، همگی تاکنون در ابزارهای بازیابی اطلاعات مورد توجه واقع شده است و نه در الگوریتم‌های دسته‌بندی. در نهایت، با توجه به اینکه خوشه‌بندی نتایج جستجو از فناوریهای نسبتاً رایج در سطوح وب بوده و طراحان موتورهای جستجو برای بهبود نتایج از این ویژگیها در بازآرایی نتایج خود استفاده نمی‌کنند، در پروژه مقایسه‌ای جهت نشان دادن این مهم صورت گرفته است.

مرجع [63]، روشی ارائه نموده که با توجه به خصوصیات متن قرآن کریم قادر است ابتدا آیات را درون موضوعاتی از پیش تعریف شده دسته‌بندی کند، سپس آیات همه سوره‌ها را با توجه به وزن کلمات آن برای هر موضوع، درون دسته‌ها طبقه‌بندی نماید. الگوریتم مورد استفاده از خانواده الگوریتم‌های بیزین بوده و مبتنی بر TFIDF می‌باشد. در این روش برای بالا بردن دقت از درخت ارتباطات و جدول کلمات هم معنی استفاده شده است. تعداد بالا داده‌های آزمایشی، یکی از نقاط قوت موجود در این پروژه بوده است.

در مقاله [64] سیستم جدیدی برای دسته بندی خودکار متون فارسی ارائه شده است. این سیستم شامل دو مرحله اصلی است: مرحله پردازش و مرحله دسته‌بندی. در مرحله اول با پردازش داده‌های آموزشی بهترین ویژگی‌های نماینده هر کلاس استخراج شده و برای آموزش دسته‌بندی کننده مبتنی بر ماشین بردار پشتیبان استفاده می‌شوند و سپس در فاز دسته‌بندی، ماشین بردار پشتیبان قادر خواهد بود داده‌های تست را به یکی از کلاس‌های آموزش داده شده نسبت دهد. در روش ارائه شده در این مقاله برای افزایش دقت دسته‌بندی کننده از دانش معنایی موجود در گنجواژه بهره گرفته شده است. به این ترتیب که ویژگی‌های مربوط به هر کلاس می‌تواند با استفاده از گنجواژه گسترش یابد.

در مقاله [65] الگوریتمی جهت دسته‌بندی نتایج جستجو مبتنی بر درخواست کاربر به منظور افزایش کارایی موتورهای جستجوگر و کاهش تلاش کاربر در یافتن نتیجه مورد نظر در لیست نتایج ارائه شده

است. ابتدا صفحه مرورگر بارگزاری شده و پیشنهادهایی برای کاربر ارائه می‌شود، برخی پارامترها مانند تعداد موتورجستجوگر برای جستجو، اولویت زبانی، تعداد دسته ها و ... به او پیشنهاد می‌گردد. پس از بازیابی درخواست کاربر که توسط واسط کاربری مطرح شده است، متن درخواست به یکی از موتورهای جستجو ارجاع داده شده و پس از دریافت نتایج مورد نظر، اسناد به دست آمده از وب پردازش شده و نتایج دسته‌بندی شده نمایش داده می‌شوند. از ترکیب وزن‌دهی TFIDF و الگوریتم K-means برای امر دسته‌بندی استفاده شده است.

مقاله [66] رویکرد دسته‌بندی معنایی مبتنی بر سیستم‌های چند عامله در بازیابی اطلاعات در وب را برای بهبود روند جستجو پیش‌رو گرفته است. ایده اولیه این روش از آنجا نشات گرفته است که بیشتر موتورهای جستجو سنتی تلاشی برای درک مفهوم درخواست ها نکرده و تنها به یافتن عینی (شمول/ عدم شمول) کلمات و عبارات می‌پردازند. بدین منظور از پایگاه هستی شناسی برای افزودن معنا و محاسبات زبان شناسی به عبارات درخواستی کاربر استفاده شده است. در این برنامه از یک سیستم شامل چندین عامل اطلاعاتی هوشمند متعامل جهت حل مسایلی که موجودیت‌های منفرد و روشهای قدیمی از پس آنها بر نمی‌آیند بهره گرفته شده است. عامل اطلاعاتی، عاملی است برای بهبود سرویس‌های جستجو و امکان دسترسی به یک یا چندین منبع اطلاعاتی که قادر به تطبیق و دستکاری اطلاعات منابع برای پاسخگویی به درخواست کاربر می‌باشد. هر کدام از متغیرهای درخواست به یک عامل داده شده و هر کدام 100 نتیجه اول را از سایت‌های پایه موجود بازیابی می-کنند. این نتایج بر اساس معیارهای مشخصی دوباره اولویت دهی می‌شوند. سپس مناسب‌ترین کلمات برای شرح بهتر مفهوم انتخاب می‌شوند. در ادامه ابعاد ماتریس حاصل به کمک روش SVD کاهش یافته و لیبل خوشه‌ها شناسایی می‌شود. سپس هر سند به لیبل مربوطه نسبت داده شده و خروجی به صورت گروه‌های مرتب‌سازی و اولویت‌بندی شده نمایش داده می‌شود. معیار ارزیابی به‌کار گرفته شده در این سیستم، میزان رضایت کاربر در معنادار بودن لیبل و تناسب لیبل با محتویات اسناد است.

برچسبزنی موضوعی که در مقاله [68] به آن پرداخته شده است، محتمل‌ترین موضوعی که محتوای متن بدان اشاره دارد را مشخص می‌کند. برای رسیدن به این هدف، از روش فاصله‌یابی در فضای بردار بسامدی بهره گرفته شده است. در این روش فاصله بین بردارهای بسامد وزن‌دار کلمات کلیدی محاسبه می‌شود. کلمات کلیدی شامل کلمات با بیشترین حاصلضرب بسامد پیکره‌ای و فاکتور وابستگی به طبقه هستند که مجموعه این کلمات از پیکره متنی زبان فارسی استخراج شده‌اند. در مرحله هرس، علاوه بر کلمات دستوری مانند حروف اضافه و ربط، کلمات کلیدی زاید نیز به‌صورت دستی بر اساس دیدگاه‌های زبان‌شناختی و معنی‌شناختی حذف می‌شوند. در گام بعد، برای هریک از طبقات مورد نظر یک بردار بسامد وزن دار کلمات کلیدی تشکیل می‌گردد. از بین این بردارها، برداری که کمترین فاصله را با بردار بسامد وزندار متن آزمون دارد، طبقهٔ مربوط به متن آزمون را مشخص می‌نماید. پایگاه داده مورد استفاده شامل 8486 کلمه در 10 دسته بوده و پایگاه کلمات زائد بیش از 60 هزار کلمه را در خود جای داده است.

از [69] به عنوان یکی دیگر از مقالاتی که به دسته‌بندی نتایج جستجو مبتنی بر درخواست کاربر به منزله سهولت کاربران در مشاهده سریع و بهینه نتایج موتورهای جستجو پرداخته‌اند، می‌توان نام برد. در این روش پس از اعلان درخواست از سمت کاربر و دریافت لیست ترتیبی اسناد از موتورهای جستجو، متد پیشنهادی به استخراج و اولویت‌دهی عبارات ممتاز به عنوان اسامی کاندید دسته‌ها بر مبنای مدل آموزشی رگرسیون خطی می‌پردازد. سپس هریک از اسناد به عبارت نزدیکتر اختصاص یافته و دسته‌های نهایی با ادغام کاندیدها ایجاد می‌گردد. الگوریتم پیشنهادی ابتدا کل عنوان و قطعه خبری بدست آمده از موتورهای جستجوگر پایه را تجزیه کرده، تمام عبارات مهم را در یک n- گرام ایجاد کرده و برای هر کدام چندین پارامتر مشخص را محاسبه می‌نماید. سپس به کمک یک مدل رگرسیون بر اساس داده‌های آموزشی قبلی و اعمال ترکیب این خصوصیات، امتیاز برجسته‌سازی عبارت را محاسبه می‌کند و تعدادی از عبارات با بالاترین امتیاز به عنوان نام و شرح دسته انتخاب می‌شوند. روش وزن‌دهی به عنوان و قطعه متفاوت است چرا که به احتمال زیاد عناوین شامل مباحث

مهم‌تری هستند. طول یا تعداد لغات تشکیل عبارت، فرکانس تکرار هر کلمه، آنتروپی دسته و استقلال عبارت جهت وزندهی و معیار تشابه cosine برای دسته‌بندی اسناد استفاده شده است.

در این بخش با تعدادی از پژوهش‌های کلی انجام شده در زمینه دسته‌بندی اسناد متنی و اسناد وب آشنا شدیم. همچنین نگاهی کلی به مقالات مرتبط با موتورهای دسته‌بند اسناد وب داشتیم. هرچند این نمونه‌ها تمام موارد پژوهشی را در بر نمی‌گیرند، اما نشان‌دهنده خط مشی و رویکردی کلی از پروژه پیش‌رو است.

فصل چهارم: روش تحقیق

1-4 مقدمه

امروزه گسترش چشمگیر تعداد وب‌سایت‌های خبری و حجم بالای اطلاعات موجود در آنها و تمایل برای کسب دقیق و به موقع اطلاعات پیرامون حوادث جاری از اهمیت بسیار بالایی برای مردم برخوردار است.

سرعت بالای گسترش اخبار در اینترنت، امکان پیگیری مداوم اخبار به صورت دستی را غیر ممکن نموده است. این حجم بالا نه تنها در بدست آوردن اطلاعات مورد نیاز به کاربران کمک نمی‌کند، بلکه باعث سردرگمی و ابهام بیشتر آن‌ها نیز می‌گردد، تا آنجا که نیاز به سیستمی برای مدیریت و در نهایت متمایز ساختن اسناد خبری از یکدیگر و دسته‌بندی آن‌ها در گروه‌های متشابه در این بستر به چشم می‌خورد. اجزای یک سیستم یادگیری نظارتی شامل عامل یادگیر، مجموعه داده‌های آموزشی و داده‌های آزمایشی(داده‌های برچسب‌دار و غیر برچسب‌دار) و تعداد مشخصی دسته می‌باشد. در کاربرد خوشه‌بندی متن، هر سند به عنوان یک داده آموزشی در نظر گرفته می‌شود. هر خوشه در این کاربرد، حاوی اسناد مربوط به یک موضوع خاص است. تعداد خوشه‌های آموزشی بر اساس تنوع موضوعی تعیین می‌گردد. این تعداد توسط کاربری که داده‌های آموزشی را فراهم می‌کند، مشخص می‌شود. برچسب داده‌های آموزشی، شماره و یا نام خوشه و یا خوشه‌هایی است که هر داده آموزشی به آن خوشه و یا خوشه‌ها تعلق دارد. عامل یادگیر در فاز آموزشی با استفاده از داده‌های آموزشی، آموزش داده شده و یک تعمیم از هر خوشه یاد می‌گیرد که عامل را قادر به شناسایی اسناد مربوط به هر خوشه می‌کند. در فاز آزمایش، عامل یادگیر، اسناد متنی را که در فاز آموزشی ندیده است و متعلق به یکی از خوشه‌های آموزشی می‌باشد، دریافت کرده و احتمال تعلق این سند را به هر کدام از خوشه-های آموزشی به دست می‌آورد. در این‌جا خوشه‌بندی برای هر سند، به معنی انتخاب خوشه‌ای با بالاترین احتمال تعلق سند به آن خوشه می‌باشد.

در این بخش راهکار پیشنهادی برای دسته‌بندی و نمایش بهینه اخبار جهت سهولت زمانی و دسترسی کاربران بیان شده است. برای رفع این معضل، از تلفیق روش‌های وب‌کاوی و دسته‌بندی داده‌های

متنی برای مدیریت بهینه اسناد خبری بهره گرفته‌شده است. قسمتی از متن اسناد خبری در ابتدا به کمک یک خدمتگزار وب از سایت‌های خبری مورد تایید و مشخص برداشته شده و پس از آموزش سیستم، پیش‌پردازش‌های مورد نیاز بر روی اسناد قابل تست صورت گرفته، هر سند خبری به دسته-های مربوطه و انتخابی کاربر ارجاع داده شده و نتیجه به صورت لیستی از اخبار دسته‌بندی شده نمایش داده می‌شود. در ادامه به معرفی و تشریح اجزای تشکیل دهنده معماری پیشنهادی می-پردازیم.

در فصل‌های قبل با مراحل روند دسته‌بندی اسناد در وب و مقدمات فنی آن آشنا شدیم. در این فصل به صورت متمرکز کار بر روی اسناد خبری حقیقی را در پیش گرفته و از بیان مجدد کلیات جلوگیری می‌کنیم.

2-4 اخبار موجود در وب

به صورت کلی وظیفه انتشار اخبار معتبر و رسمی کشور در اینترنت به عهده سه دسته سازمان می-باشد. دسته اول مربوط به نهادهای خبرگزاری کشور می‌باشد. تهیه و انتشار خبرهای رسمی کشور تنها در خبرگزاری و توسط خبرنگاران مربوط به هر واحد صورت می‌پذیرد. از جمله خبرگزاری‌های معتبر می‌توان به خبرگزاری رسمی جمهوری اسلامی ایران (ایرنا)، خبرگزاری دانشجویان ایران (ایسنا)، خبرگزاری مهر، ایلنا، ایتنا و ... اشاره نمود. نحوه دسته‌بندی اسناد خبری در این سازمان‌ها بدین گونه است که پس از ترسیم و تعریف نمودار خبری در هر سازمان (بسته به سیاست و خط مشی سازمان)، گروه‌های نهایی انتخاب شده و هر واحد، تعدادی خبرنگار مشخص و انحصاری را استخدام می‌کند. هر خبرنگار با توجه به گرایش کاری خود اخبار مربوطه را ردیابی نموده و سند خبری را تهیه می‌کند. این کار برای هر دو گروه اخبار داخلی (مراجعه مستقیم و تهیه گزارش) و اخبار خارجی (ترجمه اخبار معتبر دیگر کشورها یا تهیه گزارش حضوری) صورت گرفته و پس از تائید مسئول واحد، کارشناس خبری آن را جهت انتشار در سایت قرار می‌دهد. گزارشات مربوطه به سایر استان‌ها نیز پس از تائید فرمانداری و کارشناس خبری در گروه استانی انتشار می‌یابد. در حقیقت دسته‌بندی اسناد در

دنیای حقیقی در اولین فاز ممکن (تهیه خبر) صورت گرفته و هیچ گونه عمل جداسازی و تفکیک خبری و یا پردازش و گروه‌بندی اخبار صورت نمی‌پذیرد. قرار دادن اخبار بر روی سایت، در محدوده زمانی یا تعدادی مشخصی صورت نگرفته و هر کدام از خبرگزاری‌ها بسته به تعداد و مهارت پرسنل و بازه زمانی به صورت 24 ساعته به نشر اخبار جدید می‌پردازند. میانگین نرخ به روز رسانی اخبار سه خبرگزاری در جدول(4-1) آورده شده است.

دسته دوم که به عنوان پورتال‌های خبری کشور شناخته شده‌اند، مربوط به سازمان‌ها، نهادها و ارگان-های دولتی و خصوصی بوده و تنها اخبار درون سازمانی را برای اطلاع پرسنل و یا بخش محدودی از افراد جامعه درون وب سایت خود قرار می‌دهند. از جمله پورتال‌های سازمانی می‌توان به سایت مربوط به دانشگاه‌ها و وزارتخانه‌ها اشاره کرد. عمده اخبار انتشاری بر روی آنها شامل عزل و نسب پرسنل، بخش‌نامه‌ها و جلسات برگزارشده و تصمیمات اتخاذ شده است. بدون شک اخبار بین سازمانی و یا خبرهایی که تاثیر مستقیم بر روند امور جامعه دارند، توسط روابط عمومی سازمان و یا با تشکیل کنفرانس خبری و دعوت خبرنگاران و در برخی موارد بنا بر پیگیری خبرنگاران به اطلاع عموم مردم رسانده می‌شود.

آخرین دسته مربوط به سایت‌های خبری- تحلیلی می‌گردد که رویکردی غیر دولتی داشته و هر کدام به جریان، گروه یا فردی خاص منتسب بوده و یا مستقل عمل می‌کنند. در این دسته بیشتر با تحلیل، مقایسه و بازپخش اخبار با تاکید بر گردش و نشر آزاد افکار و اندیشه‌های مردم و تلاش در نشر دیدگاهها، نقدها و نظرهای آنان در چارچوب حفظ هویت دینی و ملی و حدود قانون و باورهای مردم مواجه هستیم. عملا کار تهیه خبر بر عهده این سازمان نبوده و هیچ دخل و تصرفی در اصل اخبار ندارند و در اکثر موارد شاهد تقسیم‌بندی‌های شخصی‌سازی شده بر اساس نظر مدیر سایت هستیم و هر کدام بر روی موضوعات خاصی تمرکز دارند. یکی از اهداف این سایت‌ها که تعداد آن‌ها نیز بسیار زیاد است جمع‌آوری مجموعه اخبار و سهولت دسترسی کاربران به مجموعه اخبار گزینشی است. از این دسته می‌توان به سایت‌های خبری تحلیلی کلمه، فردا، بامداد نیوز و ... اشاره کرد.

تمرکز و توجه پروژه کنونی بر روی اخبار انتشار یافته در خبرگزاری‌های رسمی بوده و در این میان سه خبرگزاری ایرنا، ایسنا و مهر انتخاب شده‌اند. در عین حال سیستم نهایی متعلق به هیچ کدام از گروه‌بندی‌های فوق نبوده و تنها به جمع‌آوری و گروه‌بندی اخبار دور از دسترس و نظارت کاربر انسانی می‌پردازد. به همین دلیل از عنوان موتور دسته بند اسناد خبری در بستر وب برای آن استفاده نموده-ایم.

خبرگزاری	7-6	8-7	9-8	10-9	11-10	12-11	13-12	14-13	15-14	16-15
ایرنا	13	31	23	16	26	19	14	15	17	13
ایسنا	17	23	34	26	31	25	23	27	31	30
مهر	1	18	11	9	17	8	4	6	5	12
خبرگزاری	17-16	18-17	19-18	20-19	21-20	22-21	23-22	24-23	1-24	6-1
ایرنا	10	10	11	30	18	13	15	6	6	8
ایسنا	26	26	12	17	8	8	1	0	7	3
مهر	12	6	2	8	4	3	0	1	1	2

جدول(4-1): نرخ به روز رسانی اخبار در سایت‌های پایه

3-4 پویش صفحات و جستجو لینک‌های خبری

بدون شک وجود یک پایگاه با داده‌هایی متناسب، برای هر پروژه‌ای الزامی است. از آنجا که تمرکز اصلی کار بر روی نمونه‌های به‌روز موجود در سایت‌های خبری می‌باشد، لذا ایجاد پایگاهی از داده‌های مشابه نمونه‌های حقیقی برای آموزش سیستم یکی از نخستین اقدامات است، در نتیجه باید به جستجوی اسناد خبری در سایت‌های متعلق به خبرگزاری‌های معتبر پرداخته و سعی در جمع‌آوری داده‌های اولیه نماییم. به دلیل عدم پوشانندگی تمام اخبار توسط یک سایت خبر رسانی منفرد، برای دستیابی کامل و همه جانبه به حوادث نیاز به ادغام اخبار موجود در چندین سایت خبری وجود دارد. پیشتر بیان شد که به دلیل اهمیت صحت اسناد، نمی‌توان به هر صفحه‌ای در دنیای وب مراجعه نمود، در نتیجه برای این‌کار از میان خبرگزاری‌های رسمی و مورد تایید در کشور، چند خبرگزاری را انتخاب نموده و عملیات آموزش و تست را بر روی آنها انجام دادیم. خبرگزاری‌های مورد بحث عبارتند از: خبرگزاری دانشجویان ایران(ایسنا)، خبرگزاری جمهوری اسلامی ایران (ایرنا) و همچنین خبرگزاری

مهر. بدیهی است سیستم با توجه به نیاز و درخواست کاربر قادر به گسترش و انطباق‌پذیری می‌باشد. بعد از ورود به پورتال هر یک از خبرگزاری‌ها، به استخراج بخشی از اسناد خبری می‌پردازیم.

لازم به ذکر است که بنا بر روند کار، موتورهای جستجوگر می‌توانند میتنی بر درخواست کاربر و یا بر اساس پیش‌فرض‌های تعریف شده برای سیستم عمل پویش صفحات را انجام دهند. در همین راستا اولین اقدام ما پویش صفحات خبری خاص در بازه‌های زمانی تعریف شده به منزله دسترسی به لینک خبرها است، برای این کار از یک خدمتگزار وب استفاده کرده‌ایم. بدین صورت که پس از انتخاب و تایید وب‌سایت‌های پایه، به آدرس صفحات مربوطه رفته و کد HTML ایجاد کننده صفحه را خوانده و در آرایه‌ای ذخیره می‌کنیم. پس از انجام این کار به دنبال تگ مربوط به لینک‌های درونی صفحه رفته و آدرس آنها را به عنوان مشخصه کلیدی و انحصاری هر صفحه استخراج می‌کنیم. در ادامه با ارسال یک خدمتگزار وب دیگر اطلاعات صفحات لینک را خوانده و بخش مورد نظر از هر صفحه را در قالب یک قایل به فرمت xml. ذخیره می‌کنیم.

از جمله مشکلات پیش‌رو برای افزودن خبرگزاری‌های جدید، عدم وجود استاندارد در فرمت و محتوای صفحات می‌باشد. از آنجا که هیچ گونه ساختار مشخص، قالب و همچنین زبان کدینگ واحدی برای سایت‌ها در نظر گرفته نشده و هر خبرگزاری صرفا بر مبنای تمایل و علم متخصصان درون یا برون سازمانی به صورت مجزا، اقدام به طراحی و ارائه سایت نموده، نتیجتا هنگام فراخوانی و ایجاد پایگاه نمونه‌های اولیه با تنوع بالایی از فرمت چه در سطح آرایش اخبار و چه در سطح دسته‌بندی‌های نهایی مواجه می‌شویم. نمودار خبری سه خبرگزاری مورد بحث -بدون توجه به زیر گروه‌های متعدد تعریف شده برای هر کدام در جدول (4-2) آورده شده است.

خبرگزاری	نمودارخبری
ایرنا	سیاسی، اقتصادی، اجتماعی، فرهنگی، علمی، ورزشی، خارجی، استانها، پژوهش‌های خبری
ایسنا	اندیشه امام و رهبری، علمی و فناوری، اجتماعی، دانشگاه و حوزه، اقتصادی، سیاسی، بین‌الملل، فرهنگی و هنری، معارف و حقوق، فرهنگ حماسه، ورزشی، عکس، استان‌ها
مهر	فرهنگ و هنر، فرهنگ و ادب، دین و اندیشه، حوزه و دانشگاه، فناوری‌های نوین، اجتماعی، اقتصادی، سیاسی، بین‌الملل، ورزشی، انرژی هسته‌ای، عکس، استان‌ها، دفاع مقدس

جدول (4-2): نمودار خبری مربوط به سه سازمان خبرگزاری پایه

4-3-1 مدیریت خطاهای احتمالی به کمک کدهای وضعیت HTTP

پویایی، تغییر و حذف اطلاعات در بستر نت به دلیل دسترسی همگانی به آن امری بدیهی بوده و روزانه تعداد لینک‌های تهی یا خراب بسیاری ایجاد می‌گردد. در صورت عدم وجود نظارت و مدیریت لینک‌ها احتمال ورود آدرس اشتباه به لیست لینک‌های قابل پردازش وجود دارد، که در نهایت به دلیل عدم توانایی دسترسی عامل خزنده، برنامه با خطا مواجه شده و عملکرد آن متوقف می‌گردد. از جمله خطاهایی که در مرحله تست به دفعات مکرر با آن مواجه گشتیم می‌توان به نمونه‌های زیر اشاره کرد. در صورت عدم وقوع این خطاها، کد وضعیت 200 (سالم) را به خود اختصاص داده و عملیات ارجاع به آن صورت می‌گیرد.

کد 400: درخواست بد[70]

کد 400 به دلیل درک نشدن شیوه نگارش[71] درخواست واسط کاربری از سرور رخ می‌دهد، در این حالت مفهوم تقاضای کاربر برای سرور روشن نیست و درخواست قابل پردازش نمی‌باشد، این خطا ممکن است به دلایل دیگر، از جمله نقص در انتقال داده‌ها (به فرض به دلیل قطع یا افت سرعت ارتباط) نیز رخ دهد.

[70] Bad Request
[71] Syntax

کد 403: دسترسی غیر مجاز[72]

کد 403 مربوط به مواقعی است که کاربر درخواست منبعی را از سرور دارد که دسترسی به آن برای همه کاربران محدود شده است، در اینجا حتی با ورود نام کاربری و کلمه عبور نیز امکان دسترسی مقدور نخواهد بود، معمولا مدیران سایت‌ها، دسترسی مستقیم به فولدرها و نمایش فایل‌ها به صورت لیست را غیر فعال می‌کنند، در نتیجه وقتی آدرس یک فولدر را از آن سرور درخواست می‌کنیم، با خطای 403 مواجه خواهیم شد.

کد 404: منبع درخواستی پیدا نشد[73]

کد 404 در مواقعی رخ می‌دهد که واسط کاربری تقاضای منبعی (به طور مثال یک فایل یا صفحه) را از سرور دارد که در حال حاضر موجود نبوده یا حذف شده است (و یا ممکن است نام آن تغییر کرده باشد)، البته احتمال دارد در آینده مجددا آن منبع ایجاد شده و در دسترس قرار گیرد.

کد 410: محذوف[74]

کد 410 به معنی حذف همیشگی منبع درخواستی از سرور است، بر خلاف خطای 404، کد 410 به واسط کاربری یا موتورهای جستجو می‌گوید که نباید مجددا آن منبع را درخواست کنند، چرا که برای همیشه حذف شده است، البته در عمل موارد استفاده از این کد خیلی محدود است و تنظیم خطای 404 بهتر و اصولی تر است.

کد 500: خطای داخلی سرور[75]

کد 500 به معنی وقوع یک خطای داخلی در سرور است و معمولا به دلیل نقص تنظیمات یا انجام به روزرسانی نرم افزاری یا سخت افزاری رخ می‌دهد، تنظیم این کد در مواقعی که می‌خواهیم در سایت، تغییراتی اعمال کنیم که باعث از دسترس خارج شدن آن می‌شود، می‌تواند مفید باشد.

---

[72] Forbidden
[73] Not Found
[74] Gone
[75] Internal Server Error

2-3-4 انتخاب قطعه خبری

همانگونه که بیان شد، داده‌های مورد پردازش شامل اسناد خبری انتشار یافته در اینترنت می‌باشد. یکی از مهم‌ترین تفاوت‌های موجود میان اسناد متنی اداری و اسناد متنی در وب، اندازه اسناد است، بدین‌منظور که در سیستم‌های مبتنی بر وب به دلیل اهمیت حجم و زمان پردازش‌داده تمرکز اصلی بر روی ریز اسناد بوده و پردازش مستقیم بر روی کل متن صورت نمی‌گیرد. داده‌های نهایی ما ریز اسنادی شامل: آدرس دسترسی مستقیم به خبر، شرح متن و زمان استخراج خبر می‌باشد، که در ادامه با خصیصه‌های هر کدام آشنا می‌شویم.

آدرس دسترسی مستقیم: همانطور که قبلا بیان شد، در هر صفحه تعداد زیادی لینک مشاهده می‌گردد، اما تنها لینک‌های درون‌صفحه‌ای، نشان‌دهنده ارجاع به اخبار انتشاری داشته و دیگر لینک‌ها که آنها را برون صفحه‌ای می‌نامیم، کار برقراری ارتباط میان خبرگزاری با دیگر سازمان‌ها و نهادها را بر عهده دارند. آدرس دسترسی مستقیم، به مجموع آدرس وب سایت پایه (خبرگزاری) وآدرس لینک درون صفحه‌ای که خزنده از آن استفاده نموده است گفته می‌شود. از آنجا که هر لینک تنها به یک سند خبری اشاره دارد، لذا از آن به عنوان کلمه مشخصه هر سند استفاده نموده و برای جلوگیری از ورود اسناد تکراری به پایگاه داده، پس از مشاهده متن هر کدام از اسناد، لینک دسترسی مستقیم آنرا ذخیره می‌نماییم. در انتها نیز از همین آدرس جهت رساندن کاربر به خبر مورد نظر استفاده خواهیم نمود.

شرح متن: هدف اصلی این بخش استخراج و گزینش بخشی از متن خبر برای انجام پردازش می‌باشد. جهت استخراج شرح ریزاسناد روش‌های بسیاری وجود دارد. در بسیاری از پروژه‌ها مشاهده می‌گردد که داده آموزشی و گاها داده تست توسط کاربر و به کمک پیش‌فرض‌هایی جهت سهولت بخشیدن به کار پردازش آماده می‌گردد. هر چند این عمل میزان دقت و نتیجه نهایی را تا حدی افزایش خواهد داد، اما بدون شک برای کار با داده‌ها و فضای حقیقی مشکلات بسیاری به وجود خواهد آورد. در صورتی که بخواهیم کار استخراج را به صورت دستی انجام دهیم، بدون شک مجموع متون موجود در

عنوان، توضیحات و چند جمله اول که طبیعتا چکیده کل مطلب را در بردارد، بهترین گزینه جهت فهم کلی مطلب بوده و به اندازه کافی قادر به تفکیک‌سازی عناصر مهم متون از یکدیگر می‌باشند. اما از آنجا که نمونه‌های نهایی به صورت خودکار از صفحات استخراج می‌گردد، جهت بالا بردن تطابق داده‌های تست و آموزش، در فاز آموزش نیز داده‌ها به صورت اتوماتیک توسط عامل خزنده تهیه شده-اند. یکی از روش‌های اولیه در این زمینه استفاده از نشانه‌های موجود در کد HTML اسناد می‌باشد. بدین صورت که پس از مراجعه به لینک‌های درون صفحه‌ای، متون موجود در نشانه‌های "عنوان[76]"، "توضیحات[77]" و "کلمات کلیدی[78]" را که نویسنده و یا اپراتور هنگام درج مطلب تکمیل نموده را به عنوان ریزسند انتخاب نماییم. با نگاهی به صفحه منبع سایت‌های موجود و نشانه‌گذاری‌های متفاوت (به دلیل عدم وجود استاندارد در فرمت داده‌ها) این روش کارآمد نبوده و در بسیاری موارد این نشانه-ها خالی می‌باشند. جدول (4-3) نشان‌دهنده تفاوت‌های نوشتاری در صفحه خبرگزاری‌های مورد بحث است. روش مورد استفاده در این پروژه، استخراج سه پاراگراف ابتدایی موجود در هر متن به دلیل پوشانندگی نسبی و قابل قبول محتوای خبری در تمام صفحات است.

زمان: آخرین پارامتر مهم در جمع‌آوری، انتشار و باز پخش اسناد خبری، زمان است. در سیستم طراحی شده نتایج مربوط به دسته‌بندی در آخرین زمان بازبینی اسناد نمایش داده می‌شود. از آنجا که در مورد پارامتر زمان نیز هیچ گونه فرمت مشخصی در سایت‌ها وجود ندارد، لذا زمان خبر را برابر با زمان جمع‌آوری داده از صفحات قرار داده‌ایم. زمان بازبینی صفحات توسط کاربر و یا با زمان پیش فرض 30 دقیقه تنظیم شده است.

[76] Title
[77] Description
[78] Keywords

نهایی	مهر	ایرنا	ایسنا	خبرگزاری
Base website+ inner link	inner link	inner link	inner link	<URL
-	وب سایت	* + <h1	*	<title
*	P3	<h2	Title	<p1
*	P4	P3	P1	<p2
*	P5	P4	P2	<p3
System time	-	تاریخ انتشار		<time, date
-	-	*	-	Keywords
-	-	*	* 1 sen.	Description
<p	<Span	<h1	<p1	Start point

جدول(4-3): چکیده ساختار صفحه منبع کد سایت‌های پایه و سیستم نهایی

4-4 فاز آموزش

روش دسته‌بندی اسناد در سیستم طراحی شده مبتنی بر نظارت بوده و بر آموزش تکیه دارد. به این معنی که در این روش به یک مجموعه اولیه از اسناد که قبلا تحت دسته‌هایی مشخص دسته‌بندی شده‌اند نیاز میباشد. در انتها نیز ارزیابی به کمک مجموعه آزمایشی صورت می‌پذیرد.

4-4-1 جمع آوری اسناد آموزشی

هدف این فاز جمع‌آوری قطعات متنی است که توسط موتورهای جستجو در جواب نیاز برنامه بازگردانده شده‌اند. تمام اطلاعات مورد نیاز از صفحات وب استخراج شده و تگ‌های اضافه و HTML آن جدا می‌شوند. در نتیجه تنها عنوان، محتوای قطعه ، منبع اولیه و تاریخ ثبت نگهداری می‌شود. در این قسمت صفحه‌ای مانند صفحه نتیجه جستجو گوگل با لیستی ترتیبی از اطلاعات مشابه خواهیم داشت. این اطلاعات می‌توانند بر حسب درخواست کاربر و یا بر حسب پیش فرض سیستم بازگردانده شوند که در مورد اول باید به پیش‌پردازش درخواست کاربر و اعمال قوانین جبری منطقی احتمالی تعبیه شده در صفحه رابط کاربر پرداخته شود. اما از آنجا که سیستم مورد بحث ما مبتنی بر پیش-

فرض سیستم بوده و درخواست کاربر نقشی در نوع یا تعیین کلید واژه مشخص اسناد استخراجی ندارد، در نتیجه کار را با دنبال‌گیری تمام لینک‌های خبری موجود در خبرگزاری‌ها ادامه می‌دهیم.

برای انجام نمونه‌گیری و آماده‌سازی داده‌ها از آرشیو خبری روزانه (بازه زمانی، تعداد اخبار) در دسته-های مختلف استفاده شده است. بدین صورت که هر کدام از اسناد خبری به صورت مجزا در یک فایل ذخیره شده و با علم بر طبقه مربوط به آن در پایگاه داده جمع‌آوری شده است.

به صورت عمده می‌توان چنین بیان داشت که استفاده از تعدادی سند رسمی برای آموزش این چنین سیستم‌هایی کافی به نظر می‌رسد، اما از آنجایی که اسناد مورد بررسی در زمینه اطلاعات و سازمان-های خبری از هیچ‌گونه نظم و روند مشخصی در زمان و نحوه اعلام بر خوردار نیستند، در نتیجه می-توانیم از دو متد متفاوت استفاده نماییم. در مورد اول می‌توانیم با استفاده از آرشیو خبری مربوط به بازه زمانی مشخص پروسه آموزش سیستم را انجام دهیم. اما در برخی موارد احتمال آن می‌رود که در طی بازه مورد بحث تعداد بسیار اندکی سند در رابطه با شاخه‌ای خاص وجود داشته باشد، که این امر نه تنها به روند آموزشی سیستم ضربه وارد می‌کند، بلکه میزان اعتماد و ضریب کارکرد سیستم را نیز پایین خواهد آورد. در روش دوم آموزش را با حداقل تعداد مشخص خبر از هر زیرشاخه بدون در نظر گرفتن محدودیت زمانی انجام می‌دهیم. به کمک تلفیق دو روش فوق می توان به جامعیت مورد نظر دست یافت. در این پروژه نیز از تعداد حداقل 30 سند در هر شاخه استفاده شده و اخبار در چند بازه زمانی جمع‌آوری شده‌اند. تعداد 808 سند نهایی به‌کار برده شده به تفکیک به ازا هر گروه در جدول (4-4) آورده شده است.

گروه	علمی	استان-ها	بین-الملل	اجتماعی	اقتصادی	دانشگاهی	مذهبی	فرهنگی	ورزشی	سیاسی
تعداد	78	95	78	87	82	49	57	111	84	87

جدول(4-4): تعداد اسناد آموزشی به تفکیک گروه‌ها

2-4-4 تطبیق و تعیین گروه‌ها

مشکل عدم وجود استاندارد در سایت‌های خبری در این مرحله نیز برای ما مشکل آفرین است؛ چراکه هیچ الگویی برای دسته‌های خبری وجود نداشته و هر خبرگزاری بسته به نظر شخصی مسئولان و سیاست‌های خاص سازمانی گزینش دسته‌های نهایی را انجام داده است. در این برنامه نیز مجموعه‌ای از برآیند دسته‌های خبری انتخاب شده است و در قسمت آموزش اخبار مربوط به دسته‌های متفرقه، به صورت گزینشی به دسته‌های نهایی نسبت داده شده‌اند. این مشکل در سطح بین‌المللی نیز در صنعت خبری مشاهده می‌گردد. جدول (4-5) نشان‌دهنده نمودار دسته‌های خبری مربوط به خبرگزاری‌های مورد بحث و دسته‌های نهایی انتخابی می‌باشد. پس از تکمیل فرآیند نمونه‌گیری که در بخش قبل آورده شد، اخبار به گروه‌های نهایی انتساب داده شده و جهت آموزش به سیستم، لیبل گذاری می‌شوند.

خبرگزاری	نهایی	ایسنا	ایرنا	مهر
1	علمی، فناوری	علمی و فناوری	علمی	فناوری های نوین
2	استان ها	استان ها	استان ها	استان ها
3	بین الملل	بین المللی	خارجی	بین الملل
4	اجتماعی	اجتماعی	اجتماعی	اجتماعی
5	اقتصادی	اقتصادی	اقتصادی	اقتصادی
6	حوزه، دانشگاه	دانشگاه و حوزه		حوزه و دانشگاه
7	دینی، مذهبی	معارف و حقوق		دین و اندیشه
8	فرهنگی	فرهنگی و هنری، فرهنگ حماسه، اندیشه امام و رهبری	فرهنگی	فرهنگ و هنر، فرهنگ و ادب، دفاع مقدس
9	ورزشی	ورزشی	ورزشی	ورزشی
10	سیاسی	سیاسی ، بیداری اسلامی	سیاسی	سیاسی، انرژی هسته ای
			پژوهش های خبری	

جدول(4-5): نمودار گروه بندی سازمان‌های خبرگزاری و نمودار خبری نهایی

3-4-4 آماده‌سازی اسناد آموزشی

تا این قسمت عملیات استخراج داده‌های مورد نیاز از وب‌سایت‌های پایه تعریف شده در برنامه صورت گرفته است. در ادامه به کمک برخی روش‌های مربوط به متن‌کاوی، اسناد را جهت پردازش نهایی آماده ساخته و عملیات پیش‌پردازش را بر روی آنها اعمال می‌کنیم. داده‌های آموزشی، مجموعه اطلاعاتی هستند که دانشی را به صورت پیش‌فرض به عامل یادگیر منتقل می‌کنند .در واقع تفاوت عمده دو روش خوشه‌بندی با یادگیری با نظارت یا بدون نظارت از تفاوت در نوع آموزش آنها نشأت می‌گیرد. هدف نهایی این مرحله یکسان‌سازی اسناد و حذف داده‌های ناهمگون جهت سهولت در امر پردازش است.

نرمال سازی

در دوران گذشته پیش از اختراع دستگاه‌های تحریر، هر انسان بسته به ذوق و سلیقه خود سعی در نوشتار مرتب و منظم متون با دست‌خط خویش می‌نمود. این تفاوت در شیوه‌های ظاهری نوشتاری، میزان خوانایی و درک متون توسط دیگر انسان‌ها را تا حد زیادی کاهش می‌داد. وجود تکنولوژی‌های امروزه، مشکل ناخوانایی متون را کاهش داده و نوشتار را مستقل از نویسنده نموده است. چراکه همگی کاربران با استفاده از استانداردها و کدهای تعریف شده متون را تایپ نموده و خروجی یکسانی ایجاد می‌نمایند. مشکل جدید در این زمینه وجود تعداد و تنوع بالای فونت‌ها است که به آنها قلم نگارش کامپیوتری نیز گفته می‌شود. علاوه بر شکل ظاهری متفاوت، وجود فونت‌های عربی فارسی با کدهای اختصاصی برای هر کدام و عدم پوشش برخی کاراکترها در زبان عربی، تفاوت‌های پنهان بسیاری را در مقایسه کلمات با ظاهر مشابه ایجاد نموده است. استفاده از شیوه‌های نگارشی متنوع و عدم امکان استانداردسازی نوشتاری را نیز باید به مشکلات بالا افزود. به عنوان مثال استفاده از انواع شکل همزه در میان و پایان کلمات را می‌توان نام برد.

تشخیص این تفاوت‌ها به دلیل شباهت ظاهری، برای کاربران انسانی دشوار بوده و از طرف دیگر کامپیوتر به دلیل استفاده کدهای مختلف آنها را جدا از هم و به صورت کلمات مجزا در نظر می‌گیرد.

به همین دلیل در ابتدای کار قبل از انجام هرگونه پردازشی، کد فونت‌ها را برای حروف مشابه یکسان نموده و همزه انتهایی کلمات را نیز حذف نموده‌ایم. این مرحله باید برای تمام داده‌های آموزشی، تست، کلمات اضافی، فهرست‌های مکمل و به صورت کلی برای هر متن ورودی انجام پذیرد. با این کار نه تنها افزونگی داده از بین می‌رود، بلکه قادر به ایجاد پایگاه لغات غیر تکراری نیز خواهیم شد. در جدول (4-6) کد اختصاصی به حروف مشابه و نمونه‌هایی از هر مورد آورده شده است. از آنجا که اسناد قابل پردازش، متون رسمی خبرگزاری‌ها می‌باشند، در نتیجه احتمال وجود غلط‌های املایی در نظر گرفته نشده است.

حرف	کدهای متفاوت حرف	مثال	کد نهایی
و	1572و 1608	مودب... مؤدب	1608
ه	1577و 1607و 1726و 1729	تذکره... تذکرهٔ... تذکرهٔ	1607
ی	1574و 1609و 1610و 1740	مسایل... مسائل و کتبی... کتبی	1610
ک	1603و 1705	سرکش دار عربی	1603
ء	1569	امضا ... امضاء	32

جدول(4-6): یکسان‌سازی حروف مشابه

در این مرحله همچنین کاراکترهای نگارشی به کاربرده شده در متن و فضاهای خالی میان کلمات حذف شده و با تعریف حد آستانه طول کلمه، کلمات خالص متن بدست می‌آید. در بسیاری از موارد به دلیل عدم رعایت نشانه گذاری یا فاصله، دو کلمه به هم متصل شده و عملا به عنوان داده خراب وارد متن می‌شود. از جهت دیگر طول بیشتر وندهای اضافه شده به کلمات (فعل، اسم) کمتر از حد مشخصی است. هرچند این وندها در قسمت ریشه‌یابی پس از تطبیق با جدول وندها حذف می‌گردند، اما لازمه این کار مقایسه تک تک کلمات با جدول بوده و بدون شک دارای بار محاسباتی و زمانی برای برنامه خواهد بود. لذا در این مرحله از پیش‌پردازش، تمامی لغاتی که طول کمتر از سه کاراکتر و بیش از هشت کاراکتر دارند را از مجموعه اصلی حذف می‌کنیم. با این عمل طول سند قابل پردازش به حد چشمگیری کاهش یافته و از ورود لغات با ارزش معنایی پایین جلوگیری به عمل خواهد آمد. در مورد

کاراکترهای جداساز نگارشی نیز به جای تطبیق و مقایسه کاراکترهای متنی، از کد اسکی آنها استفاده نموده‌ایم. در زیر لیستی از این کاراکترها آورده شده است.

کاراکترهای نگارشی: {" "، "|"، "/"، "."، "("، ")"، "!"، "$"، ":"، ";"، "_"، "-"، "؟"، ","، "’"، "..."، "،" }

اقدامی دیگری که در این مرحله صورت می‌گیرد، تشخیص و جداسازی کلمات فارسی است. از آنجا که کار بر روی اسناد خبری به زبان فارسی صورت می‌گیرد، تمامی کاراکترهای مربوط به دیگر زبان‌ها حذف شده و تنها، محدوده قابل پذیرش (1610> کد کاراکتر>1569) به انضمام کد کاراکترهای مخصوص زبان فارسی (1662= "پ"، 1670=" چ"، 1688=" ژ " و 1711="گ") پذیرفته می‌شود.

ریشه یابی

هدف کلی از ریشه‌یابی، تلاش برای رسیدن به بن و پایه کلمات و تبدیل صورت‌های مختلف نگارشی آنها به یک کلمه واحد است. در نتیجه این کار یافتن میزان تکرار و به تبع آن میزان اهمیت کلمات به واقعیت بسیار نزدیک‌تر خواهد بود. به عنوان مثال افعال فارسی از نظر ساختمان در یکی از دسته‌های ساده، پیشوندی، مرکب، عبارات فعلی و افعال مرکب ضمیری در سه زمان حال، گذشته و آینده برای اول، دوم، سوم شخص مجرد و جمع جای می‌گیرند. این عمل با افزودن برخی شناسه‌ها، پسوندها و پیشوندهای مشخص به بن افعال انجام می‌شود، هرچند استثناهایی نیز در تمام موارد به چشم می-خورد.

در مورد اسامی نیز با افزودن وندهای معین، کلمات را به حالت صفات مختلف، جمع تبدیل می‌گردند. یکی از مشکلات در زبان فارسی ورود و کاربرد متداول لغات و قواعد عاریه از دیگر زبان‌ها است. در این موارد به دلیل عدم وجود روال معین، امکان دسترسی به بن کلمه وجود نداشته و به عنوان کلمات استثنا شناخته می‌شوند. به عنوان مثال در مورد جمع مکسر، نیاز به ریشه‌یابی وزن عربی کلمه وجود دارد.

با انجام دادن این کار و حذف وندها از کلمات، کار حذف کلمات عمومی و افعال نیز به سادگی بیشتری صورت خواهد پذیرفت. باید توجه داشت که در موارد بسیاری با حذف وندها، به ساختار اصلی کلمه آسیب وارد شده و کلمه شکل ناقصی به خود می‌گیرد. اما از آنجایی که اولا تعداد این کلمات خیلی زیاد نیست و ثانیا عمل ریشه‌یابی بر روی تمام داده‌های ورودی به شیوه یکسان انجام می‌پذیرد، در نتیجه تاثیر منفی آن قابل اغماض و چشم‌پوشی است.

حذف کلمات عمومی و افعال

انواع کلمه در زبان فارسی عبارتند از: فعل، اسم، صفت، ضمیر، قید، حرف و شبه جمله، که در این میان قیود، ضمایر و حروف تنها برای شکل‌دهی و بالا بردن فهم جملات کاربرد داشته و به صورت مستقل حامل پیام خاصی نیستند. لذا تعداد تکرار این کلمات در اسناد بسیار بالابوده اما نقش خاصی در تمایز و طبقه بندی اسناد ندارند. در بخش قبل با انجام عمل ریشه‌یابی تمامی کلمات را به شکل واحد بن آنها درآورده و در این قسمت به تشخیص و حذف کلمات عمومی (مستقل از کاربرد) می‌پردازیم. در اکثر زبان‌ها لیستی از کلمات عمومی پر تکرار وجود دارد، که در مرحله پیش پردازش آنها را حذف می‌نماییم. عملیات حذف در دو مرحله قبل و بعد از عمل ریشه‌یابی صورت گرفته و نقش بسزایی در کاهش تعداد کلمات قابل پردازش دارد. در این قسمت تمام ضمایر، انواع حروف ربط و اضافه، بن افعال و برخی کلمات پر تکرار عمومی حذف می‌گردند.

تقسیم بندی

هدف از این کار تشخیص محدوده لغات قابل پردازش در یک متن است. استفاده از علائم نشانه‌گذاری در این قسمت کاربرد فراوانی دارد. اما از آنجا که در قسمت قبل تمام کاراکترهای نشانه‌گذار و جداساز کلمات تعیین و حذف شده‌اند، در نتیجه از کد مربوط به کاراکتر "فضای خالی" جهت تعیین محدوده کلمات استفاده نموده‌ایم. در صورتی‌که متن اصلی استخراجی به صورت خام وارد برنامه میشد، می‌توانستیم برای جداسازی ریز سند نیز از کدهای موجود در `html` استفاده نماییم. اما در این پروژه جداسازی متن ریز سند در اولین مرحله صورت گرفته است.

آماده‌سازی اسناد آموزشی، علاوه بر یکسان‌سازی نوشتاری لغات، باعث کاهش 32.6 درصدی، معادل 33132 کلمه پرارزش قابل استفاده در بخش بعد شده است. این کاهش تعداد لغات، تسریع زمان و قدرت پردازش را به همراه داشته و بالطبع کارایی سیستم را بهبود داده است.

4-4-4 ایجاد پایگاه لغت

پس از تکمیل فرآیند پیش پردازش، متن اسناد به لغات کلیدی تشکیل دهنده آنها تقسیم‌بندی شده و جهت آموزش سیستم تمام لغات وارد یک پایگاه متمرکز می‌شوند. بر خلاف بسیاری از پروژه‌های داده‌کاوی امکان قرار دادن حد آستانه برای لغات با مرتبه تکرار بسیار بالا و یا بسیار پایین وجود ندارد. مهم‌ترین دلیل این امر پویایی اسناد خبری است. به عنوان مثال امر انتخابات ریاست جمهوری هر چهار سال یکبار در ایران رخ می‌دهد. در صورتی که لغات مربوط به اولین سند منتشره در مورد انتخابات، به دلیل تعداد پایین دفعات تکرار در محدوده لغات قابل پذیرش قرار نگیرند، بدون شک هیچ‌گاه این خبر وارد پایگاه داده نخواهد شد، لذا در این سیستم هر کلمه به صورت منفرد ارزش داشته و از ورود هیچ یک از کلمات به پایگاه داده کلی نمی‌توان صرف‌نظر کرد.

1-4-4-4 ادغام پایگاه‌های لغت کمکی

در قسمت قبل بیان شد که هدف کلی از ساخت دیکشنری، ایجاد یک پایگاه داده منظم و دسته-بندی‌شده از اسناد آموزشی جهت تطبیق، مقایسه و تعیین گروه نهایی اسناد تست می‌باشد. این امر به صورت دریافت اسناد مجزا از گروه‌بندی‌های مختلف در سایت‌های پایه بوده و پس از پاکسازی، مجموعه لغات هر گروه در آرایه‌ای متناظر گردآوری شده‌اند. به کمک این روش به جای مقایسه هر سند ورودی با کل اسناد، هر سند را با مجموعه لغات نهایی تطابق می‌دهیم. از آنجایی که تعداد اسناد آموزشی محدود می‌باشد (از نظر تعداد و زمان)، در نتیجه امکان پوشش کلی دایره لغات مربوط به هر دسته با تکیه بر مجموعه آموزشی امکان‌پذیر نمی‌باشد.

راهکاری که برای رفع این مشکل در نظر گرفته شده است، استفاده از فهرست لغات مکمل می‌باشد. در این فهرست سعی بر آن شده است که تا حد توانایی لغات کلیدی متمایز کننده مربوط به هر دسته

شناسایی شده و به صورت یک پایگاه کمکی در اختیار سیستم قرار داده شود. به عنوان مثال در دسته استان‌ها، نام تمامی شهرهای ایران، در بخش بین‌الملل، نام کشورهای مختلف، در قسمت حوزه و دانشگاه، لیستی از نام تمامی دانشگاه‌ها و رشته‌های تحصیلی و در بخش هنری سبک‌های مختلف موسیقی، شعر و نام نویسندگان و شعرا و آلات موسیقی آورده شده است. در دیگر بخش‌ها نیز تلاش شده است تا تخصصی ترین کلمات موجود اضافه گردند. بدون شک این فهرست تمامی لغات را در بر نداشته و هیچ گونه ادعایی بر این موضوع نخواهیم داشت. یکی از مهم‌ترین این دلایل پویایی زبان و حجم بالای لغات جدید و اصطلاحات و ترجمه‌های گوناگون از کلمات دیگر زبان‌ها می‌باشد.

در تهیه فهرست تکمیلی باید توجه داشت که لغات تا بالاترین حد ممکن ویژگی تمایزی گروه‌ها را پوشش دهند. به عنوان مثال در نام‌گذاری دانشگاه‌ها یکی از مرسوم‌ترین روش‌ها اختصاص نام شهر یا استان در ادامه نوع دانشگاه می‌باشد. حال آنکه از این لغات در فهرست استان‌ها به عنوان نمونه‌ای انحصاری نام برده شده است. در نتیجه در معرفی دانشگاه‌ها اسامی خاص دیگر گروه‌ها (مانند نام شهر، استان و کلماتی چون امام، شهید و ...) حذف شده است. برای ایجاد این فهرست‌ها سعی در استفاده از آخرین نسخه‌های ویرایش شده در دانشنامه آزاد ویکی پدیا نموده‌ایم.

2-4-4-4 تعیین و حذف کلمات پرکاربرد

هر چند در ابتدا تعدادی از لغات به عنوان کلمات عمومی و پرکاربرد از مجموعه اصلی هر سند حذف شدند، اما به عنوان آخرین اقدام در فاز آموزش سیستم، کار حذف کلمات پرکاربرد مربوط به حوزه خبرگزاری را انجام می‌دهیم. بدین منظور با تحلیل لیست لغات پایگاه‌های گروهی، کلماتی را که در تمام گروه‌ها به دفعات مکرر استفاده شده‌اند را ، فارغ از توجه به معنی و ارزش کلمه، از پایگاه کلمات حذف نموده و به لیست کلمات عمومی، جهت حذف شدن آنها در فاز تست، اضافه می‌کنیم. از جمله می‌توان به کلماتی همچون خبر، خبرگزاری، بلامانع، ذکر و نام خبرگزاری‌ها اشاره کرد. هرچند این لغات بار معنایی منحصر به فرد دارند، اما تکرار بیش از اندازه آنها، اهمیت متمایز کنندگی‌شان را در

این پروژه از بین برده است. مجموعه این لغات به انتهای لیست کلمات عمومی اضافه شده و در فاز تست از این لیست نهایی به جای لیست اولیه استفاده شده است.

3-4-4-4 تعلق کلمات به گروه‌ها

در این بخش علاوه بر ایجاد پایگاه داده اصلی که پیشتر به آن پرداختیم، به تعداد گروه‌های نهایی نیز پایگاه منفرد ایجاد می‌گردد. اما تفاوت عمده پایگاه لغت گروهی با پایگاه اصلی، کلمات تشکیل دهنده آنها است. همانطور که بیان شد، در پایگاه اصلی هر کلمه به محض مشاهده برای اولین بار به لیست اضافه شده است، این عمل تنها وجود یا عدم وجود کلمه را در بردارد، اما در مورد گروه‌ها، تمام لغات موجود در پایگاه اصلی به انضمام تعداد تکرار هر کدام در دسته مربوطه آورده شده است. بدون شک، عدم وجود یک کلمه در گروه، مقدار عددی صفر را در آرایه متناظر در بر خواهد داشت.

در نتیجه این قسمت، ماتریسی با ابعاد [تعداد دسته، تعداد کل لغات غیر تکراری] ایجاد می‌گردد. در ادامه از این ماتریس جهت میزان تعلق لغات برای گروه‌بندی استفاده خواهیم نمود. در اکثر پروژه‌های دیگر از ماتریسی مشابه برای اسناد استفاده می‌شود. به منزله یافتن میزان تعلق کلمه به گروه، از روش احتمال پیشین[79] فرمول دسته‌بندی بیزین در که فصل دوم توضیح داده شد، با اندکی تغییرات جهت بهینه‌سازی جواب استفاده نموده‌ایم.

برای این منظور با توجه به گسسته بودن مقادیر برای محاسبه $P(X_k|C_i)$ از فرمول (4-1) استفاده می‌کنیم‌که در آن احتمال تعلق لغت kام (X_k) به دسته i ام (C_i) بدست می‌آید. تعداد کل دسته‌ها در پروژه ده دسته تعریف شده است. در نتیجه این عمل پارامتری به نام (NGF) به ازا، هر لغت در هر گروه بدست می‌آید.

$$P(X_k|C_i) = \frac{C_i \text{ فرکانس تکرار هر کلمه برای موضوع}}{\text{فرکانس تکرار هر کلمه برای کل موضوعات}} \qquad (1-4)$$

[79] Prior Probability

5-4 فاز تست

عملیات دسته‌بندی اسناد خبری به صورت دستی، بدین شکل انجام می‌گیرد که هر خبرنگار، بسته به نوع گرایش خبری که دارد، هر یک از گزارشات خبری مربوط به حوزه خود را تهیه و تدوین نموده و برای انتشار در اختیار مسئول واحد مربوطه قرار می‌دهد. از آنجا که این افراد قبلاً توسط متخصصان آموزش داده شده‌اند، اشراف کلی بر روی موضوعات داشته و بالطبع هیچ گونه مشکلی برای دسته-بندی نهایی وجود نخواهد داشت.

اما در ادامه باید هر کدام از اسناد خبری آزمایشی در حوزه‌های تعریف شده پوشش داده شوند. در نتیجه پس از اتمام فاز آموزش، هر سند به یک یا بیش از یک گروه خبری نسبت داده می‌شود.

1-5-4 تهیه قطعات

در این قسمت به تهیه اسناد خبری برای تست سیستم خواهیم پرداخت. یکی از مسایل مهم در روند اجرای برنامه، تلاش برای جلوگیری از ورود و ذخیره اطلاعات مشابه و تکراری است، چرا که سیستم بر اساس آموزه‌هایش اسناد تکراری را به درستی دسته‌بندی نموده و این کار میزان اعتمادپذیری و کارایی آن را در مورد داده‌های حقیقی کاهش خواهد داد. به همین منظور آدرس تمام صفحات پیمایش شده، ذخیره شده و در صورت عدم تشابه، صفحه مربوط به لینک خبری جدید به عنوان نمونه تست پذیرفته می‌شود. داده‌هایی که به منظور آزمایش سیستم پیشنهادی مورد استفاده قرار گرفته‌اند، برگرفته از مجموعه اخبار انتشار یافته در سایت‌های پایه با در نظر گرفتن حداقل 10 سند خبری در هر گروه، در چندین بازه زمانی متفاوت است که توسط سیستم پویشگر از سایت‌های پایه بدست آمده، تعداد اسناد آموزشی نهایی مورد استفاده در هر گروه جهت ارزیابی سیستم در جدول (7-4) آورده شده است.

گروه	علمی	استانی	خارجی	اجتماعی	اقتصادی	دانشگاهی	مذهبی	فرهنگی	ورزشی	سیاسی	رسانه-های دیگر	پژوهش خبری
تست	22	30	57	27	27	18	14	49	51	56	20	1

جدول(4-7): تعداد اسناد آزمایشی به تفکیک گروه‌ها

2-5-4 آماده‌سازی اسناد آزمایشی

در این مرحله تمام داده‌های تست جهت یکسان شدن با اسناد آموزشی و سهولت در امر پردازش، مانند قسمت قبلی آماده شده و عملیات نرمال سازی، ریشه یابی، تعیین و حذف کلمات و افعال عمومی و تقسیم بندی متن به کلمات قابل پردازش با تمام قوانین و شرایط ذکر شده پیشین بر روی آنها اعمال می‌گردد. در نهایت به دلیل آماده‌سازی مشابه گروه آموزشی و آزمایشی، به داده‌هایی با فرمت و ظاهر مشابه دست خواهیم یافت. از آنجا که در نهایت متن اصلی خبر به انضمام دسته نهایی برای پذیرش و اعتبارسنجی به کاربر نمایش داده خواهد شد، لذا باید متن اولیه بدون اعمال دیگر پیش‌پردازش‌ها ذخیره گردد.

3-5-4 وزندهی به کلمات

پس از اتمام مرحله پیش پردازش و آماده سازی داده‌ها، به تعداد حداقل کلمات با مفهوم درون یک متن دست یافتیم که می‌تواند بیانگر خلاصه متن بوده و به ما در درک کلی مفهوم متن یاری رساند. عمل وزن‌دهی به کلمات به منزله یافتن میزان اهمیت و نقش کلمه در عبارات و اسناد صورت می-پذیرد. بدون شک کلمه با تعداد تکرار زیاد نسبت به طول متن، نشانگر تعلق متن به دسته خاص خواهد بود. این تعلق با مقایسه وزن کلمات سند با فرکانس تکرار کلمات مشابه، در پایگاه داده متعلق به گروه‌های نهایی که در قسمت آموزش ایجاد شده بودند، مشخص می‌گردد. برای این کار ابتدا لیستی شامل تمام لغات تشکیل دهنده متن و فرکانس تکرار آنها در سند اولیه بدست می‌آید. برای نرمال نمودن مقادیر بدست آمده، بعد از شمارش تعداد تکرار به ازا هر کلمه، حاصل را بر تعداد کل کلمات تشکیل دهنده متن تقسیم می‌نماییم. با انجام این عمل پارامتر (TF) برای هر لغت موجود در متن تست بدست خواهد آمد. در ادامه با مقایسه هر کلمه با کلمات موجود در پایگاه لغات اصلی، احتمال حضور کلمات در هر یک از گروه‌ها را استخراج نموده و به ماتریس قبلی اضافه می‌کنیم. در نتیجه این عمل ماتریسی در ابعاد [تعداد کلمات کاندید سند تست×12] خواهیم داشت که ستون اول و دوم به ترتیب مربوط به لغات کاندید متن تست و تعداد تکرار نرمال شده آنها و 10 ستون دیگر

متعلق به احتمال حضور کلمه در هر گروه است. سپس به محاسبه میزان تعلق سند به گروه‌ها به کمک فرمول (2-4) می‌پردازیم:

$$Similarity(D,G) = \sum_{w=1}^{\#\ of\ D\ words} \sum_{g=1}^{10} \frac{TF_w \times TF_w NGF_g}{\sum_{k=1}^{10} TF_w NGF_k - TF_w NGF_g} \qquad (2-4)$$

در فرمول فوق w معادل تعداد کلمات غیر تکراری سند آزمایشی پس از اعمال پیش‌پردازش‌ها و g و k بیانگر تعداد دسته‌ها است. در مخرج سعی بر آن شده است که تا حد امکان، میزان تمایز نتایج نهایی با ارجاع چندگانه به پارامترهای اصلی، افزایش یابد. TF_w فرکانس تکرار کلمه در متن و $TF_wNGF_{g/k}$ فرکانس تکرار در گروه‌ها را نشان می‌دهد. در نهایت وزن هر سند (D) در هر گروه (G) بدست خواهد آمد. نتایج در نهایت به صورت درصد تشابه نمایش داده می‌شوند.

4-5-4 انتساب اسناد و تعیین مقدار آستانه

بعد از یافتن میزان تعلق سند به هر دسته، بدون شک گروه با بالاترین مقدار تشابه، به عنوان گروه برنده معرفی شده و لیبل مربوط به سند را به خود اختصاص خواهد داد. در موارد بسیاری شاهد آن هستیم که یک سند خبری به بیش از یک گروه خبری اختصاص دارد، دلیل این امر همپوشانی موضوعات خبری و عدم تفکیک جنبه‌های مختلف رخدادهای روزمره زندگی است. به عنوان مثال حادثه مربوط به آتش‌سوزی در یکی از مدارس شهرستان‌های کشور، علاوه بر اینکه یک خبر مربوط به استان‌ها می‌باشد، اما در مقیاس کشوری به مسایل اجتماعی، سیاسی نیز بستگی دارد. همچنین نقد و همدردی ادبی مربوط به این حادثه در دسته فرهنگی نیز جای خواهد داشت. به همین دلیل نمی‌توان در تمام موارد تنها گروه با بالاترین میزان تشابه را به عنوان گروه برنده معرفی نمود، بلکه با در نظر گرفتن مقدار آستانه، گروه‌های با تشابه بیش از مقدار مشخص را به عنوان گروه‌های برنده با درجه تشابه متفاوت نشانه‌گذاری خواهیم کرد. روش مورد استفاده برای تعیین گروه برنده، استفاده از الگوریتم k- نزدیک‌ترین همسایه است که در این مورد نزدیک بودن به معنای مشابه بودن متن ریز سند و ریز سند ایجاد شده به کمک کلمات تشکیل دهنده متن تست در هر دسته است. از آنجا که انتخاب عدد k وابسته به متن و ارزش کلمات آن در گروه‌های مجزا است، لذا انتخاب یک عدد ثابت

روش بهینه‌ای محسوب نمی‌شود. به همین دلیل به جای انتخاب عدد معین، تعداد گروه‌ها تا زمان رسیدن به حداقل میزان تشابه را با یک آستانه‌گذاری مشخص نموده‌ایم. تعیین مقدار آستانه در اغلب موارد به صورت تجربی و پس از تحلیل نتایج به ازا هر مقدار اولیه انجام می‌گردد. در این پروژه نیز از ترکیب روش تجربی تحلیل نتایج فنی و درک انسانی استفاده شده است. بدون شک قرار دادن یک عدد ثابت به عنوان حد آستانه نتیجه درستی در پی نخواهد داشت، چرا که تمام اسناد از توزیع یکسانی برای تعلق به دسته‌ها استفاده نمی‌کنند. تفاوت در تعداد و محتوای لغات مورد استفاده در متون نیز به این دلیل دامنه می‌زند. فرمول(4-3) به محاسبه حد آستانه در محدوده صفر (عدم تطابق کامل) و بیش از 100 درصد (دسته‌بندی نرم)، به ازا هر سند به صورت اختصاصی می‌زند.

$$Threshold = \frac{\max(Similarity) + \min(Similarity) + 4 \times \frac{\max(Similarity) + \min(Similarity)}{2}}{4} + \min(Similarity) \quad (3-4)$$

اساس و پایه این روش بر مبنای برآورد سه نقطه‌ای[80] تخمین زمان و هزینه در مدیریت پروژه‌ها می‌باشد که در این شرایط شما باید یک برآورد در شرایط خوش‌بینانه، یک برآورد در شرایط بدبینانه و یک برآورد در شرایط عادی داشته باشید. در این حالت با یک فرمولی که شامل وزن‌دهی به این شرایط می‌باشد، می‌توان زمان و هزینه فعالیت‌های پروژه را تخمین زد. این فرمول معمولا به این صورت است که برآورد فعالیت برابر است با مجموع برآوردهای خوشبینانه و بدبینانه به علاوه چهار برابر برآورد در شرایط نرمال تقسیم بر مجموع ضرایب به کار رفته در صورت. در این پروژه به ترتیب حداکثر و حداقل درصد تعلق سند به گروه را معادل با شرایط خوش‌بینانه و بدبینانه در نظر گرفته و میانگین زمانی را شرایط عادی می‌نامیم. بدین صورت فرمول به شکل بالا درآمده است. در نهایت با جمع نمودن مجدد حداقل مقدار، نتیجه بدست آمده را با میزان درک کاربر انسانی تطابق داده‌ایم. در برخی موارد مشاهده شده است که به دلیل نزدیکی میان درصد تشابه گروه‌ها هیچ دسته ای مقدار بیش از حد آستانه را ندارد که در این مواقع از بالاترین نسبت با چشم‌پوشی از مقدار آستانه استفاده

[80] Three point Estimates

می‌کنیم. یکی از موارد قابل بحث در این بخش تعیین گروه‌بندی پایه و مرجع به منزله مقایسه نتایج خروجی سیستم با آن است. هرچند روند آموزش سیستم بر مبنای گروه‌بندی اخبار در وب‌سایت‌های پایه بوده است، اما به دلیل وجود تعداد تناقض زیاد بین دسته‌های معرفی‌شده و دسته‌های حقیقی و وجود ریزمتن‌های خبری تست متعلق به دسته‌های تعریف نشده اولیه در سیستم (در قسمت نتیجه-گیری به تفصیل بیان شده است)، گروه‌بندی پایه دیگری اضافه نموده‌ایم. در این گروه‌بندی جدید داده‌های تست توسط یک کاربر انسانی نشانه‌گذاری شده و عمل مقایسه و ارزیابی سیستم بر مبنای آن صورت گرفته است. جدول (4-8) میزان تشابه/ عدم تشابه گروه‌های خبرگزاری، انتسابی (سیستم) و انتخابی (کاربر) را در مورد داده‌های تست نشان می‌دهد.

میزان تشابه		خبرگزاری- انتسابی		انتسابی- انتخابی		خبرگزاری- انتخابی		هر سه گروه	
عدم تشابه	0	%21	77	%10	36	%16	56	%2	8
متشابه	حداقل	%79	44	%90	61	%84	46	%74	48
	حداکثر		230		254		249		211

جدول (4-8) تشابه/ عدم تشابه گروه‌های خبرگزاری، انتسابی و انتخابی

در مورد گروه انتخابی از جانب کاربر انسانی، قطعات خبری بدست آمده به ده گروه 37 تایی به انضمام نام گروههای نهایی در اختیار 10 داوطلب قرار داده شده است. به منزله انتخاب سند از هر کدام از افراد خواسته شد که ریز متن را مطالعه نموده و بنا به سلیقه و درک شخصی سند را با حداقل یک و حداکثر تعداد مورد تائیدی خود نشانه‌گذاری نمایند. در هر دسته خبر سعی شده است تا تنوع گروه خبری لحاظ گردد. افراد تکمیل کننده فرم‌ها در بازه سنی 24 الی 60 سال با حداکثر تمایز در دیدگاه، علایق و سن، به منزله پوشش حداکثری دیدگاه‌های متفاوت در آن انتخاب شده‌اند.

در جدول (4-9) نتایج بدست آمده از سیستم برای چندین نمونه آزمایشی به همراه گروه‌بندی اصلی سند خبری و گروه انتسابی از جانب کاربر انسانی آورده شده است.

شماره خبر	متن ریز خبر دریافتی	عنوان گروه اصلی	عنوان گروه انتسابی سیستم	عنوان انتخابی کاربر
1	به گزارش خبرنگار اقتصادی ایرنا ، اردشیر سعد محمدی روز دوشنبه در نخستین نشست مطبوعاتی با اصحاب رسانه در مسند شرکت یادشده ،اظهار داشت تامین نیاز فولاد کشور یک ضرورت است و دستیابی به تولید میلیون تن فولاد در دستور کار قرار دارد وی با بیان اینکه تولید فولاد کشور در سال گذشته به میلیون تن رسید ، گفت ظرفیت ما در این بخش اکنون میلیون تن است و انتظار می رود تا پایان سال جاری بیش از میلیون تن تولید داشته باشیم و رسیدن به رقم میلیون تن تولید نیز پیش بینی می شود سعد محمدی خاطرنشان ساخت آمارعملکرد تولید هشت ماهه فولاد حاکی از رشد درصد نسبت به مدت مشابه سال گذشته است و میلیون و یکصد هزار تن تولید داشتیم وی، طرحهای فولادی را قائنات ، چهارمحال وبختیاری ، بافت ، سبزوار ، شادگان ، میانه و نی ریز ذکر کرد و افزود ...	اقتصادی	اقتصادی	اقتصادی
2	رییس دستگاه دیپلماسی بامداد دوشنبه در گفت وگوی اختصاصی با خبرنگار ایرنا در باره اظهارات اخیر وزیر امور خارجه آمریکا مبنی بر تمایل واشنگتن برای مذاکره مستقیم با تهران افزود مذاکره موضوعی در موارد خاص تاکنون با آمریکا شده که از آن جمله مذاکره در مورد افغانستان و عراق است وی افزود همچنین مذاکره در باره پرونده هسته ای نیز در قالب مذاکرات ایران با گروه + که آمریکا نیز در آن حضور دارد، انجام می شود، اما اگر در اذهان عمومی مذاکره فراگیر سیاسی مطرح شده باشد، این مساله در حدود اختیارات رهبر معظم انقلاب است و معظم له تشخیص می دهند که مثلا چنین اقدامی شود یا نشود وزیر امور خارجه یادآور شد مذاکره موضوعی تاکنون در چند مرحله شده است ...	سیاسی	سیاسی	سیاسی خارجی
3	به نوشته این منابع خبری ، کشتی های حامل سامانه های موشکی پاتریوت آلمان ، امروز بندر ' تراومونده ' در استان ' لوبک ' آلمان را به سوی بندر 'اسکندرون' ترکیه ترک کردند براساس این گزارش ، دو سامانه موشکی پاتریوت آلمان روز ژانویه دوم بهمن برای استقرار در ترکیه وارد بندر اسکندرون خواهند شد به گزارش منابع خبری ترکیه ، یک هیات نظامی متشکل از تن از نظامیان هلند و نفر از نظامیان آلمان نیز جهت فراهم کردن زمینه برای استقرار سیستم های موشکی ...	خارجی	سیاسی خارجی	سیاسی خارجی
3	به گزارش ایرنا، 'سسیل دوفلو ' روز دوشنبه در باره نامه درخواست کمک خود از کلیسا، گفت از اسقف اعظمی پاریس خواسته ام 'ساختمان های تقریبا' خالی در اختیار خود را برای اسکان بی خانمانها، افراد و خانواده های فاقد مسکن مناسب در دسترس این وزارتخانه قرار دهد وی تاکید کرد که تصرف و تملک ساختمان های خالی عمومی و در اختیار شخصیت های معنوی، ازهم اکنون تا پایان سالجاری میلادی محقق خواهد شد دوفلو با بیان اینکه امیدواراست دراین زمینه نیازی به اعمال قانون و قدرت از سوی دولت نباشد، افزود این قابل درک نخواهد بود اگر کلیسا نخواهد، برای محقق شدن اهداف و برنامه دولت در زمینه اسکان بی سرپناهان، همراهی و همیاری نکند دوفلو طرحی را به شورای وزیران فرانسه ارایه کرده است که براساس ...	خارجی	سیاسی	خارجی
4	به گزارش ایرنا از سازمان لیگ حرفه ای فوتبال، تیم فوتبال دسته یکی برق شیراز نیز از حضور در جام حذفی آزاد سازی خرمشهر انصراف داد بر اساس این گزارش، پیش از این نیز تیم آلومینیوم هرمزگان، شهرداری تبریز، سنگ آهن بافق و نفت مسجد سلیمان نیز از حضور در جام حذفی انصراف داده بودند بدین ترتیب مسابقات ...	ورزشی	ورزشی	ورزشی
5	به گزارش خبرگزاری مهر، دانشمندان تا کنون به طور کامل نمی دانستند چگونه حیوانات رنگها را می بینند چرا که رنگدانه های بصری در چشم های آنها شامل کروموفورهای کاملا شبیه به هم (بخش جذب نور مولکولها) هستند اما با این حال می توانند طول موج های مختلف نور را جذب کنند.همه حیوانات دارای کروموفور شبکیه ای (آلدئید ویتامین آ) هستند و بسته به پروتئین های گیرنده نور (اوپسین ها) مرتبط با این کروموفورها، همان مولکول می تواند طیفی از رنگها را از آبی و حتی فرابنفش تا قرمز جذب کند.اینکه چگونه یک تک مولکول می تواند این کار را انجام دهند تا کنون یک راز بود ...	علمی	علمی	علمی
6	برنامه 24 ساعته برای درمان سرماخوردگی.تصور کنید که یک روز از خواب بلند می‌شوید و احساس می‌کنید سردرد دارید، بدنتان درد می‌کند و بینی‌تان گرفته است ابتلا به سرماخوردگی اولین حدسی است که می‌زنید با این حال کارهای زیادی است که باید انجام دهید در این شرایط چه تصمیمی می‌گیرید آیا با همان کسالت و درد لباستان را می‌پوشید و به محل کار می‌روید یا تصمیم می‌گیرید یک ...	رسانه دیگر	اجتماعی	اجتماعی

جدول(4-9): نمونه ریز متون تست و دسته‌های نسبت داده شده به آنها

4-5-5 بازسازی مجدد پایگاه لغات

پس از انتساب سند به گروه یا گروه‌های مربوطه، باید عملیات بازسازی مجدد پایگاه لغات و وزندهی کلمات در گروه‌ها صورت پذیرد. این روش به آموزش و یادگیری سیستم در حین عملکرد کمک بسیاری خواهد نمود. در موارد زیادی مشاهده می‌گردد که به عنوان مثال با یک رویداد سیاسی یا اقتصادی جدید مواجه می‌شویم که بالطبع اسامی خاص و تازه‌ای نیز در خود خواهند داشت. با این دامنه لغات گروه با کلمات خاص تعریف شده آموزش می‌بیند و در آینده با ضریب تشابه بالاتری اسناد مشابه را دسته‌بندی خواهد کرد. در این قسمت پس از مقایسه کلمه با پایگاه لغات اصلی در صورت غیر تکراری بودن کلمه اضافه شده و یک واحد به گروه/ گروه‌های مربوطه اضافه می‌شود و نتایج کار مستقیما برای سند تست بعدی قابل استفاده است. یکی دیگر از دلایل عدم تمایل به ورود سند تکراری در مرحله تست نیز همین قسمت است.

4-6 نمای کلی برنامه

آخرین مرحله برنامه، به نمایش اسناد خبری در قالب دسته‌های نهایی و میزان تعلق خبر به دسته‌ها دارد. هر کاربر در لحظه آغاز قادر به گزینش یک یا چندین شاخه مورد نظر می‌باشد. این عمل به منزله شخصی سازی و تطابق برنامه با نیازها و علاقه مندی‌های کاربر صورت گرفته است. در انتهای کار، متن خبری کامل (متن قبل از ریشه‌یابی و حذف کلمات تکراری و عمومی) به همراه لیبل متعلق به گروه‌های نشانه‌گذاری شده و درصد تعلق به آنها با فرمت مورد نظر در یک فایل متنی ذخیره می-گردد. استفاده از متن خام و اولیه در خوانایی و تطابق نتیجه با نظر کاربر، افزایش جذابیت و وضوح مفهوم تاثیر بسیاری دارد. در ابتدا قسمت جمع‌آوری و پویش اسناد خبری به کمک محیط برنامه نویسی سی‌شارپ انجام می‌گیرد، در ادامه پردازش متن در محیط مطلب صورت گرفته و در نهایت مجددا به کمک برنامه سی‌شارپ نتایج در رابط کاربری طراحی شده نمایش داده می‌شود. تصویر (4-1) نشانگر نمای کلی برنامه و خروجی نهایی است.

شکل(4-1): نمای کلی برنامه

4-7 ارزیابی

تا این قسمت به معرفی داده‌های مورد استفاده و روند آموزش، تست و عملکرد سیستم پرداختیم. حال به بررسی و تحلیل نتایج بدست آمده و ارزیابی سیستم خواهیم پرداخت. اولین و متداول‌ترین روش برای تعیین میزان کارایی الگوریتم، معیارهای دقت(P)، فراخوانی(R)، دقیق بودن(A) و معیار شباهت F است. هرچند این معیارها در فصل قبل توضیح داده شده‌اند، اما در این قسمت نیز به صورت خلاصه فرمول هرکدام آورده شده است. برای محاسبه این معیارها به چهار پارامتر کلی نیاز داریم که در جدول (4-10) آورده شده‌اند.

متن	نسبت داده شده به گروه G_g	نسبت داده نشده به گروه G_g
متعلق به گروه G_g	True Positive (TP)	False Negative(FN)
متعلق به غیر گروه G_g	False Positive (FP)	True Negative (TN)

جدول(4-10): پارامترهای مربوط به معیارهای ارزیابی

معیار دقیق بودن، معمول‌ترین معیار مورد استفاده برای تعیین میزان دقت در الگوریتم‌های یادگیری ماشین بوده و به کمک رابطه (4-4) محاسبه می‌گردد:

$$A = \frac{TP + TN}{TP + TN + FP + FN} \qquad (4-4)$$

معیار دقت، میزان دقیق بودن یک جستجو را به کمک رابطه (4-5) بدست می‌آورد:

$$P = \frac{TP}{TP + FP} \qquad (5-4)$$

و در نهایت معیار فراخوانی، میزان کامل بودن یک جستجو را از طریق فرمول (4-6) اندازه‌گیری می‌کند.

$$R = \frac{TP}{TP + FN} \qquad (6-4)$$

در روش آخر نیز با ترکیب دو مقدار دقت و فراخوانی، معیار تشابه F (4-7) را بدست می‌آوریم:

$$F - measure = \frac{2 * R * P}{R + P} \quad (7-4)$$

نتایج حاصل از ارزیابی سیستم نسبت به گروه اصلی به ازا 372 سند آزمایشی از سه وب‌سایت پایه نسبت به گروه‌بندی انتسابی و انتخابی در جدول (4-11) آورده شده است. با توجه به اطلاعات فوق مقادیر زیر به دست آمده است.

	گروه انتسابی نسبت به گروه خبرگزاری				گروه انتسابی نسبت به گروه انتخابی			
پارامتر / گروه	دقیق بودن	دقت	فراخوانی	معیار F	دقیق بودن	دقت	فراخوانی	معیار F
علمی، فناوری	98.00	94.11	72.72	82.05	98.38	100	75	85.71
استان ها	92.30	61.53	26.66	37.21	96.77	71.42	55.55	62.5
بین الملل	92.87	75.80	82.45	78.99	93.27	93.84	74.39	82.99
اجتماعی	92.31	50	77.77	60.87	95.16	82.97	79.59	81.25
اقتصادی	95.73	73.07	70.37	71.69	97.84	88.88	82.75	85.71
حوزه، دانشگاه	99.43	90	100	94.73	99.73	100	95.23	97.56
دینی، مذهبی	97.43	85.71	42.85	57.14	97.58	100	40	57.14
فرهنگی	93.73	75.47	81.63	78.43	95.96	81.66	92.45	86.72
ورزشی	98.57	92.59	98.03	95.23	99.46	100	96.42	98.18
سیاسی	82.33	47.11	87.5	61.25	87.90	64.86	92.30	76.19
میانگین	94.27	74.54	74.00	71.76	96.20	88.36	78.37	81.39

جدول (4-11): نتایج حاصل از ارزیابی سیستم نسبت به گروه انتخابی و خبرگزاری

همانگونه که پیشتر نیز بیان شد، معیار دقیق بودن یک جستجو بیان کننده مقدار داده متعلق به گروه غیر دسته مربوطه است، که الگوریتم از ورود آنها ممانعت به عمل آورده است. در نتیجه پایین بودن این عدد نشان می‌دهد که تعداد زیادی از داده‌های متعلق به دسته‌های دیگر در این دسته جای می-گیرند. با مشاهده جدول فوق نفوذ همه جانبه اطلاعات در دسته سیاسی در هر دو گروه‌بندی مشاهده می‌گیرد، که این امر با توجه به ذات داده‌های مورد پردازش دور از ذهن نیست.

معیار فراخوانی نشان دهنده میزان اسنادی است که به صورت صحیح در دسته مربوطه نشانه‌گذاری شده و به دسته دیگری ارسال نشده‌اند. بالا بودن این معیار نشان‌دهنده میزان تفکیک‌پذیری دیگر دسته‌ها از دسته مربوطه است. این مسئله به وضوح در مقولات دانشگاهی و ورزشی نشان داده شده است.

با وجود تمام مشکلات میانگین میزان دقت دسته‌بندی بالای 90 درصد در هر دو دیدگاه، مهر تائیدی بر نحوه عملکرد صحیح الگوریتم انتخابی برای نوع داده‌های ورودی مورد بحث است.

یکی از دیگر مسائل مهم در ارزیابی سیستم، رعایت سه جزء کلی پوشش اخبار در وب، نحوه بروز-رسانی و شاخص‌گذاری اخبار به منزله نمایش آنها به کاربر است. سازمان‌های خبرگزاری پایه به نحوی انتخاب شده‌اند، که پوشش حداکثری اخبار مورد تائید در کشور را دارا بوده و با توجه به اینکه به-صورت متوسط (با توجه به جدول 1-4) در هر ساعت 41 سند خبری است، در نظر گرفتن زمان معادل 60 دقیقه برای به استخراج داده‌های جدید منطقی به نظر می‌رسد. در هنگام نمایش نتیجه نیز، اسناد بنا بر سیلقه و انتخاب کاربر به ترتیب در گروه‌های مجزا ارائه می‌شوند.

فصل پنجم: نتیجه‌گیری و پیشنهادات

1-5 مقدمه

متن‌کاوی اسناد وب به عنوان یک رشته مشترک مابین داده کاوی و یادگیری ماشین قرار گرفته است، که سعی دارد با تکیه بر الگوریتم‌ها و ترکیب متدهای این دو فیلد علمی به استخراج، کشف و مدیریت این دسته از داده‌ها بپردازد. در این طرح پژوهشی با ادغام روش‌های موجود و گزینش روش بهینه بر مبنای داده‌های اولیه به تعریف سیستم دسته‌بندی اسناد خبری در بستر وب پرداخته شده است. بدین‌منظور پس از گزینش دو دسته نهایی، کار آموزش سیستم را با تعداد 808 سند آموزشی از وب-سایت مربوط به سه خبرگزاری رسمی کشور آغاز نموده و پس از پاک‌سازی و آماده‌سازی داده‌ها به انتخاب کلمات کلیدی آنها پرداخته و پایگاه آموزشی نهایی را ایجاد نموده‌ایم. در ادامه عمل تست سیستم را با 372 داده ادامه داده و در نهایت کارایی سیستم را بر مبنای عنوان بیان شده در خبرگزاری و گروه انتخابی کاربران انسانی محاسبه نموده‌ایم. لازم به ذکر است که تمامی داده‌های مورد استفاده به کمک یک خدمتگذار وب، از منبع صفحات خبری استخراج شده‌اند. در این سیستم بیشترین تاکید بر روی بالا بردن ارتباط میان کلمات کلیدی انتخابی و نفش آنها در دسته‌های مختلف است. در نهایت k نزدیک‌ترین همسایه با در نظر گرفتن مقدار آستانه حداقل تشابه به عنوان گروه‌های انتسابی سیستم معرفی شده و به انضمام درصد تعلق نمایش داده شده است.

پیاده‌سازی و ارزیابی سیستم حاکی از دقت بالای 90 درصد انتساب سند به گروه‌های خبری بر مبنای گروه پیشنهادی خبرگزاری پایه و گروه انتخابی کاربران است. در ادامه به بیان برخی موارد که تاثیر منفی در روند آموزش، تست و ارزیابی سیستم داشته‌اند پرداخته و در نهایت اقدامات مورد نظر برای ادامه کار بر روی سیستم پیشنهادی بیان شده است.

2-5 تحلیل نتایج

در این قسمت با مشاهده نتایج بدست آمده از سیستم و تحلیل گروه‌های انتسابی به مواردی که در کارایی سیستم اخلال ایجاد نموده و همچنین موارد خاص استخراجی مانند لزوم ایجاد دسته‌های جدید و ادغام دسته‌های موجود اشاره نموده‌ایم. مباحث پایه تمامی این موارد در فصل قبل مطرح شده و از دوباره‌گویی و کپی مجدد جداول و نتایج در این بخش صرف‌نظر شده است.

عدم تطابق عناوین نهایی با نمودارهای خبری:

یکی از مشکلاتی که پیشتر نیز در مورد آن صحبت نمودیم، عدم تطابق و هماهنگی یک به یک میان عناوین نهایی گروه‌های انتخابی در برنامه و عناوین مربوط به دسته‌ها تعریف شده در نمودار خبری سازمان‌های خبرگزاری است. به عنوان مثال در خبرگزاری "ایرنا" اخبار مربوط به موضوع مذهبی به عنوان زیرشاخه‌ای در گروه فرهنگی جای داده شده است. تاثیر این امر، بالطبع منجر به ایجاد اشتباه در فرآیند آموزش گشته و در ادامه داده‌های تست با عنوان غیر مذهبی نشانه‌گذاری می‌شوند. در مورد اخبار دانشگاهی نیز در این سازمان هیچ‌گونه گروه یا زیرگروه مشخصی در نظر گرفته نشده است، حال آنکه صرف‌نظر از این دو دسته خبری در گروه‌بندی نهایی امکان‌پذیر نبوده و با جامعیت کار تضاد دارد. هرچند این عمل از دید مسئولین خبری بلامانع است اما در ارزیابی سیستم، به عنوان یک خطا محسوب می‌گردد. از دیگر نمونه‌ها در خبرگزاری "ایرنا" می‌توان به بخش "پژوهش‌های خبری" اشاره نمود که به دلیل عدم همخوانی با دیگر داده‌ها در بخش آموزش حذف گشته است، اما در داده‌های تست مشاهده شد. این مطلب در جدول (4-7) نیز قابل مشاهده است.

وجود دسته‌های از پیش تعریف نشده:

در قسمت تست با برخی داده‌ها با عنوان "رسانه‌های دیگر" مواجه گشتیم که در قسمت آموزش و همچنین در نمودار اولیه خبرگزاری "ایسنا" نامی از آن برده نشده بود. از آنجا که فرآیند آموزش بر اساس گروه‌های تعریف شده صورت گرفته است، لذا این اسناد خبری متعلق به هیچ گروهی نبوده و در نهایت امکان ارزیابی را با مشکل مواجه نموده است. در واقع در نهایت امر تعدادی داده نشانه

گذاری شده، بدون نشانه اصلی وجود دارد که تاثیر منفی بر نتایج بدست آمده وارد کرده است. جدول (4-7) حاکی از آن است که تعداد 20 سند خبری استخراجی تحت این عنوان نشانه‌گذاری شده‌اند، که بدلیل عدم وجود در دسته‌های نهایی در ارزیابی به عنوان داده‌های محسوب می‌گردند که به اشتباه نشانه‌گذاری شده‌اند. به همین دلیل این داده‌ها در ارزیابی کارایی سیستم در نظر گرفته نشده‌اند.

عدم دسته‌بندی صحیح اولیه:

در ابتدای فصل نحوه جمع‌آوری و نشانه‌گذاری سنتی اسناد در سازمان‌های خبرگزاری را بیان نمودیم. در هر سازمان، گروه‌بندی بسته به حوزه خبری و کاری هر خبرنگار معین گشته و به صورت معمول اخبار در بخش مربوطه انتشار می‌یابند. در موارد متعددی مشاهده شده است که خبر نه تنها به گروه تعلق ندارد، بلکه در زمینه انتشار خبر نیز کار به درستی صورت نگرفته است. به عنوان مثال تصویر (5-1) یک سند خبری است که برخلاف تهیه آن توسط خبرنگار علمی، به عنوان یک خبر اقتصادی منتشر شده است. از این زمینه ناهمگونی‌ها به شدت در صفحه اصلی سایت‌های پایه مشاهده می‌گردد که بی‌شک توجه بیشتر مسئولین امر را می‌طلبد.

شکل(5-1): نمونه ای از تناقض در گروه بندی پیشنهادی خبرگزاری‌ها

همراستایی گروه‌ها:

با مطالعه و مشاهده نتایج بدست آمده، می‌توان چنین بیان داشت که مذهب در فرهنگ مردم و جامعه تا جایی نفوذ پیدا کرده است که امکان تفکیک این دو مورد از یکدیگر به سختی حاصل می‌شود. لذا شاید بتوان مذهب را به عنوان زیر شاخه گروه فرهنگی در نظر گرفت. از طرف دیگر عنوانی مانند "فرهنگ حماسه" هرچند در دید اول بیشتر به مقولات فرهنگی می‌پردازد، اما از آنجا که بیشترین عنصر در این زمینه فرهنگ جنگ و دفاع مقدس و پیامدهای آن می‌باشد، لذا اکثر اخبار در این زمینه همراستای اخبار و رویدادهای سیاسی کشور شکل گرفته است. سیستم نیز در نهایت تمام اخبار به دسته فرهنگ حماسه را به صورت ترکیبی فرهنگی و سیاسی نشانه‌گذاری کرده است.

در همین زمینه می‌توان به اخبار ارسالی از مراکز استان‌ها و یا شهرستان‌ها نیز اشاره کرد. به عنوان مثال سخنرانی یکی از سیاستمداران یا اختصاص بودجه و افتتاح یک طرح عمرانی در یک شهر، با وجود تعلق خبر به شهر مربوطه بدون شک در دید کشوری در دسته اخبار سیاسی اقتصادی نیز قرار می‌گیرند. در موارد بسیاری نیز مشاهده می‌گردد که به عنوان مثال ارزش فرهنگی ساخت یک فیلم یا مستند از محل فیلمبرداری بیشتر بوده و نام شهر موردنظر در مقایسه با فعالیت اصلی، قابل چشم-پوشی است. این در حالی است که در وب‌سایت مربوط به خبرگزاری‌ها هر عنوان تنها به یک دسته تعلق داشته و این امر نیز در ارزیابی نهایی، سیستم را با مشکل مواجه نموده است.

از دیگر مواردی که پس از آزمایش تست مشاهده شد، می‌توان به گروه‌بندی جدید "سیاست خارجه" نام برد. در بسیاری موارد متن خبری، متن ترجمه و یا ارسالی از جانب خبرگزاران خارج از مرز بوده و در نهایت در دسته خارجی قرار داده شده‌اند. اما این اخبار در عین حال بر مسائل سیاسی متمرکز بوده و از آنجا که سیاست خارجه و علی‌الخصوص اخبار خاورمیانه تاثیر مستقیم بر سیاست داخلی ایران دارد، لذا قرار گرفتن آنها در بخش سیاسی نیز الزامی است. در همین راستا سیستم در حدود 29 مورد را با هر دو عنوان سیاسی خارجی نامگذاری کرده و مهمتر از آن اینکه از نقطه نظر کاربر انسانی نیز این نامگذاری دسته در بیش از 22 مورد مشاهده شد.

ردپای سیاست نیز (همانند روند معمول و متداول جامعه) در هر واقعه و رخدادی قابل مشاهده و پیگیری بوده و تمام جوانب کار کشور را تحت تاثیر قرار داده است، این امر در نحوه عمکرد سیستم نیز به وضوح مشاهده می‌گردد.

3-5 اقدامات آینده

بدون شک هدف نهایی انجام این پروژه ، فراتر از معرفی یک سیستم دسته‌بندی اسناد بوده و سیستم طراحی شده، تنها به عنوان سیستم پایه برای اهدافی است که در آینده قصد دستیابی به آنها را داریم. آخرین بخش این گزارش به بیان اقدامات آینده در راستای تکمیل پروژه می‌پردازد.

گستردگی داده‌های مورد پردازش:

در بخش قبل مشاهده شد که تنها از اطلاعات موجود در سه سازمان خبرگزاری در سیستم استفاده شد، هرچند این سه سازمان به دلیل اهمیت آنها در کشور، حجم بالایی از اخبار را پوشش می‌دهند، اما در ادامه قصد وارد نمودن دیگر سازمان‌های خبرگزاری به منزله دستیابی کامل به تمام موارد خبری جاری در کشور را داریم. این مسئله به منزله استفاده از وب‌سایت‌های تحلیلی خبری و یا پورتال‌های سازمانی و یا استفاده از موتورهای جستجوگر برای جمع آوری اخبار موجود در وبلاگ‌ها و سایر منابع خبری غیر رسمی نیست. چرا که با این کار نه تنها حجم اطلاعات غیر مفید بالا می‌رود، بلکه به دلیل عدم نظارت مراجع ذیربط بر کار این دسته صفحات امکان ورود اخبار غیر واقعی افزایش می‌یابد.

ادغام داده‌های خبری:

استفاده از چندین پایگاه خبری، علاوه بر پوشش حداکثری وقایع، مارا با حجم عظیمی از اسناد تکراری مواجه خواهد نمود. بیان این اطلاعات مشابه، نه تنها کاربر را بار دیگر دچار سردرگمی می‌کند، بلکه باعث به هدر رفتن زمان و فضای اختصاص یافته برای اخبار می‌گردد. به همین دلیل یکی از اقدامات مورد نظر تلاش به منزله ادغام اسناد خبری مشابه و ذکر یک خبر به انضمام لینک دسترسی به اخبار مشابه آن در دیگر وب‌سایت‌ها است. از آنجا که هر خبرگزاری بلافاصله پس از وقوع یک

حادثه سعی در آماده سازی و انتشار خبر دارد، در نتیجه در هر بار به روزرسانی سیستم (بازبینی صفحات) به احتمال زیاد با اخبار مشابه مواجه خواهیم شد. به همین دلیل ادغام این اخبار در درجه اول پیش از انتساب آنها به گروه‌ها صورت می‌گیرد. این امر در بالا بردن نقش کلمه در سند و همچنین کاهش تعداد کلمات قابل مقایسه با دسته‌های دیگر و نهایتا کاهش زمان دسته‌بندی اسناد نقش بسزایی خواهد داشت.

ایجاد جریانات خبری:

یکی از مواردی که این پروژه بر مبنای آن طراحی شده است، عدم استفاده از درخواست کاربر به منزله نمایش اسناد خبری است. یعنی برخلاف پروژه‌های بسیاری که در آن پس از اعلام درخواست کاربر با حجم محدود اطلاعات به منزله دسته‌بندی مواجه هستند، این سیستم مستقل از کاربر بوده و تلاش در پوشش تمام اخبار انتشار یافته را دارد. تنها نقش کاربر انتخاب دسته‌هایی است که تمایل دارد اخبار مربوط به آن‌ها را مرور نماید. تا این قسمت از کار به بیان اخبار داغ در هر بازه بازبینی صفحات پرداخته‌ایم. خبر داغ به خبری گفته می‌شود که اخیرا به وقوع پیوسته و انتشار یافته شده است. گام بعدی برقراری ارتباط معنادار میان اخبار جدید و آرشیو خبری است. بدین صورت که اخبار مشابه با یکدیگر در هر گروه با یک عنوان جدید (به کمک الگوریتم‌هایی که در فصل دوم به آنها اشاره شد) نامگذاری شده و به کاربر نمایش داده شود. ایجاد این جریانات خبری به مطالعه روند یک واقعه پرداخته و تاریخچه‌ای غیر مدون از واقعه را در اختیار کاربر قرار خواهد داد.

مراجع

[1] Nasraoui O., Soliman M., Saka E., Badia A.and Germain R., (2008), "A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, vol. 20, 1041-4347.

[2] Achananuparp P., Han H., Nasraoui O. and Johnson R., (2007), "Semantically Enhanced User Modeling", Proceedings of the 2007 ACM Symposium on Applied Computing, SAC '07. ACM, New York, NY, 1335-1339.

[3] Tan A.-H. (1999), "Text mining: The state of the art and the challenges", In Proc of the Pacific Asia Conf on Knowledge Discovery and Data Mining PAKDD'99 workshop on Knowledge Discovery from Advanced Databases, pages 65-70.

[4] صفی خانی ژ. و عبدالله زاده بارفروش ا.، (1384) ، "طراحی SPAM: ابزاری برای کاوش در شبکهٔ گستردهٔ جهانی"، یازدهمین کنفرانس سالانه انجمن کامپیوتر ایران.

[5] قادریان م.، (1387) ، پایان نامه ارشد: "بهبود مدل کاربر در وب‌سایت بصورت خودکار با استفاده از معناشناسی با مفاهیم خاص دامنه"، دانشکدهٔ کامپیوتر، دانشگاه صنعتی امیر کبیر، تهران.

[6] مطهری نژاد ح. ر. ، (1382) ، پایان‌نامه کارشناسی‌ارشد: "بازیابی کارا و مؤثر اطلاعات وب از طریق دستاوردهای یادگیری ماشین: طراحی و تکامل روشهای یادگیری تقویتی در کاوش متمرکز"، ، دانشکدهٔ کامپیوتر، دانشگاه صنعتی امیر کبیر، تهران.

[7] LookOFF E-book,(2002), Engine Basice, http://www.LookOff.com/tactics/engines.php3.

[8] Uramoto, N., Matsuzawa, H., Nagano, T., Murakami, A.,Takeuchi, H. and Takeda, K., (2004), "A text-mining system for knowledge discovery from biomedical documents" IBM Systems Journal., Vol.43: 516-533.

[9] Eick, C., F., Zeidat, N. and Zhao, Z., (2004), "Supervised Clustering-Algorithms and Benefits" 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'04): 774-776.

[10] Finley, T. and Joachims, T., (2005), “Supervised clustering with support vector machines” ICML '05 Proceedings of the 22nd international conference on Machine learning.

[11] Berkhin, P., (2002), "Survey Of Clustering Data Mining Techniques" Springer. Vol. 12: 1-56.

[12] Osinski S., (2004), “Dimensionality Reduction Techniques For Search Results Clustering”, Msc.thesis,computer science department of sheffield university, UK.

[13] Osinski S., Stefanowski J., Weiss D., (2004), “Lingo: Search Results Clustering Algorithm Based on Singular Value Decomposition”, Springer, 359--368.

[14] Osi´nski S. and Weiss D., (2005), “A Concept-Driven Algorithm for Clustering Search Results”, IEEE Computer Society, Intelligent Systems.

[15] Wu, X., Kumar, V., Quinlan, J., R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G., J., Ng, A., Liu, B., Yu, P., S., Zhou, Z., H., Steinbach, M., Hand, D., J. and Steinberg, D., (2007), "Top 10 algorithms in data mining", ICDM, Springer-Verlag London Limited. Vol. 14.

[16] ANDERBERG, M. R. (1973), “Cluster Analysis for Applications”, Academic Press, Inc., New York, NY.

[17] JAIN A.K., MURTY M.N., FLYNN P.J., (1999), “Data Clustering: A Review”, ACM Computing Surveys, Vol. 31, No. 3.

[18] Zamir O. and Etzioni O., (1998), “Web Document Clustering: A Feasibility Demonstration”, ACM New York, NY, USA.

[19] Dumais S. T. et.al, (1988), “Using Latent Semantic Analysis to Improve Access to Textual Information”, Proceedings of the Conference on Human Factors in Computing Systems.

[20] Carpineto C., Osin´ ski S., Romano G., Weiss D., (2009), “A Survey of Web Clustering Engines”, ACM Computing Surveys (CSUR).

[21] Pavlol, D., Balasubramanyan, R., Dom, B., Kapur, S. and Parikh, J., (2004), "Document preprocessing for naive Bayes classification and clustering with

mixture of multinomials" Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining.

[22] Isa, D., Lee, L., H., Kallimani, V., P. and RajKumar, R., (2008), "Text Document Preprocessing with the Bayes Formula for Classification Using the Support Vector Machine", IEEE Trans, Knowledge and Data Engineering. Vol. 20: 1264-1272.

[23] guyon, I. and Elisseff, A., (2003), "An Introduction to variable and feature selection", Machine learning research., Vol. 3: 1157-1182.

[24] Yuan, R., Li, Z., Guan, X. and Xu, L., (2008), "An SVM-based machine learning method for accurate internet traffic classification", Springer. Vol. 12: 149-156.

[25] Joachims, T., (1998), ‘‘Text Categorization with Support Vector Machines: Learning with Many Relevant Features’’, In European Conference on Machine Learning (ECML).

[26] حقیری م.، (1390) ، پایان نامه ارشد: "استفاده از تکنیک خوشه بندی با نظارت برای دسته بندی نامه های دریافتی در سامانه 137 شهرداری تهران"، دانشگاه شیراز.

[27] Qi, Y., Zhang. Y. and Song M., (2009), "Text Mining for Bioinformatics: State of the Art Review" iccsit, 2nd IEEE International Conference on Computer Science and Information Technology: 398- 401.

[28]احمدی م. ح. و منجمی ا. ح. و آیت س.، (1389) ، ‘‘دسته‌بندی متون فارسی با استفاده از قوانین انجمنی"، چهارمین کنفرانس دادهکاوی ایران، تهران :دانشگاه .صنعتی شریف.

[29] Han, J. and Kamber, M., (2006), "Data Mining: Concepts and Techniques" 2nd ed, SanFrancisco: Morgan Kaufmann.

[30] محمدی جنقرا م. و آنالوئی م.، (1386) ، "استخراج کلمات کلیدی اسناد فارسی"، سیزدهمین کنفرانس ملی انجمن کامپیوتر ایران.

[31] Taghva, K., Beckley, R., Sadeh, M., (2005), ‘‘A Stemming Algorithm for the Farsi Language’’, ITCC '05 Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'05) - Volume I - Volume 01 Pages 158-162

[32] پورمعصوی آ. و کامیار ح. و کامیار م.، (1387) ، "گزارش مطالعه و بررسی الگوریتم‌های موجود برای ریشه یابی کلمات"، دانشگاه فردوسی مشهد.

[33] Rahimtoroghi E., Faili H., Shakery A., (2010), ‘‘A Structural Rule-based Stemmer for Persian’’, International Symposium on Telecommunications.

[34] صفائی ع. و بدیع ک. و عباسیان ع.، (1387) ، "یک ریشه یاب برای واژگان عاریتی زبان فارسی"، چهاردهمین کنفرانس ملی سالانه انجمن کامپیوتر ایران، دانشگاه صنعتی امیرکبیر.

[35] Porter M.F., (1980), "An Algorithm for suffix stripping". Program, 14(3):130t137.

[36] Sheykh Esmaili K., Rostami A., (2006), "List of Persian Stopwords", Technical Report No. 2006-03, Semantic Web Research Laboratory, Sharif University of Technology, Tehran, Iran, June.

[37] Taghva K., Beckley R., Sadeh M., (2003), ‘‘A List of Farsi Stopwords’’, Technical Report, Information Science Research Institute, University of Nevada, Las Vegas.

[38] Davarpanah M.R., Sanji M., Aramideh M., (2009) "Farsi lexical analysis and stop word list", Library Hi Tech, Vol. 27 Iss: 3, pp.435 – 449.

[39] مقدم چرکری ن. و زمانیان م.، (1388) ، "دستهبندی متون فارسی با استفاده از الگوریتمSVM و بررسی روشهای کاهش خصیصه"، سومین کنفرانس دادهکاوی ایران، تهران :دانشگاه علم و صنعت ایران.

[40] Yang, Y. and Pedersen, J. O., (1997), "A Comparative Study on Feature Selection in Text Categorization", Proc. ICML: 412-420.

[41] ملکی م. و عبدالله زاده بارفروش ا.، (1385) ، "TFCRF : روش جدید وزن دهی ویژگی مبتنی بر اطلاعات کلاس در حوزه طبقه بندی مستندات"، دوازدهمین کنفرانس بی نالمللی انجمن کامپیوتر ایران.

[42] Salton, G., Yang, C.S., (1973), ‘‘On the Specification of Term Values in Automatic Indexing’’, Journal of Documentation, Vol. 29, No. 4, pp. 351-357.

[43] Lan, M., Sung, S.Y., Low, H.B., Tan, C.L., (2005), ‘‘A Comparative Study on Term Weighting Schemes for Text Categorization’’, IEEE International Conference on Neural Networks (IJCNN05), pp. 546-551.

[44] عربی نرئی س. و وحیدی اصل م. و مینایی بیدگلی ب.، (1386) ، "استخراج کلمات کلیدی جهت طبقه بندی متون فارسی"، دانشگاه صنعتی امیرکبیر.

[45] Shamsfard M., Hesabi A., Fadaie H., Famian A., Bagherbeigi S., Mansoory N., Fekri E., Monshizadeh M., Assi S.M., (2010), "**Semi-Automatic** Development of FarsNet; The Persian WordNet", Accepted at Global WordNet Conference (GWA2010), India.

[46] Hassanpour H., Darvishi A., (1387), "A New Similarity Measure For Signal Processing Applications", IEEE International Workshop on Signal Processing and Its Applications.

[47] Sánchez M. M., Gutiérrez E. B., Gallinas R. B., García A. M. F., Vidales M. A. S., (2008), "Clustering of Web Documents: Full-text or Snippets?", IADIS.

[48] Ferragina P., Gull A., (2004), "The Anatomy of a Hierarchical Clustering Engine for Web-page, News and Book Snippets", Data Mining, Fourth IEEE International Conference.

[49] Sayyadi H., Salehi S., AbolHassani H., (2006), "NeSReC: A News meta-Search Engines Result Clustering Tool", Advances and Innovations in Systems, Computing Sciences and Software Engineering.

[50] Zamir, O. AND Etzioni, O. (1999), "Grouper: A dynamic clustering interface to Web search results", Comput. Netw. 31, 11–16, 1361–1374.

[51] Zhang,D., Dong,Y., (2004), "Semantic, hierarchical, online clustering of Web search results". In Proceedings of 6th Asia-PacificWeb Conference (APWeb). Lecture Notes in Computer Science, vol. 3007. Springer, 69–78.

[52] Giannotti, F., Nanni, M., Pedreschi, D., Samaritani, F. (2003), "WebCat: Automatic categorization of Web search results". In Proceedings of the 11th Italian Symposium on Advanced Database Systems (SEBD), 507–518.

[53] Ferragina, P., Gulli, A. (2004), "The Anatomy of SnakeT: A Hierarchical Clustering Engine for Web- Page Snippets", In Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases. Lecture Notes in Computer Science, vol. 3202. Springer, 506–508.

[54] Subhashini R., Jawahar Senthil Kumar V., (2010), "The Anatomy of Web Search Result Clustering and Search Engines" Indian Journal of Computer Science and Engineering Vol. 1 No. 4 392-401.

[55] Lancaster, F.W., and Fayen, E.G. (1973). "Information Retrieval On-Line" Los Angeles, CA: Melville Publishing Co. Chapter 6.

[56] Ezzati Jivan, N., Fazeli, M., Yousefi, K., (2010), "New Approach for Automated Categorizing and Finding Similarities in Online Persian News", 2nd international symposium on information management in a changing world.

[57] Lee S., Kim H., (2008), "News Keyword Extraction for Topic Tracking", Fourth International Conference on Networked Computing and Advanced Information Management, NCM.

[58] Wang C., Zhang M., Ma S., Ru L., (2008), "Automatic Online News Issue Construction in Web Environment", ACM, (IW3C2).

[59] Farhoodi M., Yari A., (2010), "Applying Machine Learning Algorithms for Automatic Persian Text Classification", Advanced Information Management and Service (IMS).

[60] Li Z., Zhou D., Juan Y., Han J., (2010), "Keyword Extraction for Social Snippets", WWW '10 Proceedings of the 19th international conference on World wide web, Raleigh, North Carolina, USA, 1143-1144.

[61] نعمتی ش. و بصیری م.، (1386) ، "دستهبندی اسناد فارسی با استفاده از الگوریتم KNN"، همایش منطقهای برق و الکترونیک باشگاه پژوهشگران جوان واحد فسا.

[62] رحمتی م. و رضوی ر.، (1382) ، "یک روش سریع برای خوشه بندی نتایج جستجوی وب"، اولین کنفرانس بین المللی فناوری اطلاعات و دانش، تهران، دانشگاه صنعتی امیرکبیر.

[63] ارسطوپور ش.، (1388) ، "دسته بندی نتایج جستجو بر مبنای ویژگیهای مدارک و امکان سنجی استفاده از الگوریتمهای خوشه بندی مختلف در سطح وب". فصلنامه کتابداری و اطلاع رسانی. شماره 46. صفحه 145 – 174.

[64] دقاقزاده ح. و عبداللهی ازگمی م. و دقاقزاده م.، (1388) ، "ارائه روشی بر پایه کاوش معنایی به منظور بالا بردن دقت نتایج جستجو در نرمافزارهای قرآنی"، سومین کنفرانس دادهکاوی ایران، تهران، دانشگاه علم و صنعت ایران.

[65] مقصودی ن. و همایون پور م. م.، (1388) ، "ارائه روشی جدید در طبقه‌بندی متون فارسی با استفاده از دانش معنایی"، پانزدهمین کنفرانس بین المللی سالانه انجمن کامپیوتر ایران.

[66] Lipai A., (2008), ‘‘Web Metasearch Result Clustering System’’, Revista Informatica Economica 4(48).

[67] Bassma S. Alsulami, Maysoon F. Abulkhair, Fathy A. Essa, (2012), ‘‘Semantic Clustering Approach Based Multi-agent System for Information Retrieval on Web’’, IJCSNS International Journal of Computer Science and Network Security, VOL.12 No.1.

[68] عبدی قویدل ه. و وزیرنژاد ب. و بحرانی م.، (1389) ، "برچسبزنی موضوعی متون فارسی"، دانشگاه صنعتی شریف، تهران، ایران.

[69] Zeng H., He Q., Chen Z., Ma W., Ma J., (2004), ‘‘Learning to Cluster Web Search Results’’, 27th annual international ACM SIGIR conference on Research and development in information retrieval, ACM New York, NY, USA.

Abstract

Nowadays the need and tendency of being announced about news, has an unavoidably and direct effect on making decision in personal, social and political fields. On the other hand, Uncontrolled growth of website numbers and their information has becomes one of the most important problems which cause many difficulties for users on the World Wide Web.
By the way, wide speed of unsupervised news publishing and spreading on the web, and impressive increase of news websites in different fields, has reduced the ability for checking them. This huge amount of information has confused users instead of helping them to achieve their purpose. Because of it, design and implementation of an automatic news clustering system for managing the news on the web has explainable. Proposed method tries to resolve the problem by using data and web mining methods for clustering text news snippets.
At the first, a web client has defined for crawling the valid news agencies for acquiring news for learn and test levels. The clustered news in pre-defined groups has been showed as the result of the project. User can select both base news agency and beloved groups for final representation.
At last, too many researches have been done till now for clustering texts but all of these efforts use either term frequency or document frequency for weighing extracted keywords and then tried to make a comparison between different similarity measures, but recently a few numbers of researches have been focused on clustered features instead of documents and worlds. In this thesis we intend to use group's features by building a semi text with main text words in each group and calculating its similarity. The proposed architecture consists of a number of components namely news extraction, preprocessing, group database creation, keyword extraction, keyword

weighting, clustering, and database reconstruction. Our main contribution is that we introduce a mechanism to automatically catch and cluster news snippets on the web. At last, the groups which have passed the defined threshold are labeled as winner groups. Our evaluations show that the proposed method along with news snippet structures represents it’s efficiently and effectiveness.

Keywords*: Clustering, News Documents, Data Mining, Web Mining*

Shahrood University of Technology

Faculty of Information Technology &
Computer Engineering

An Automatic Method for Clustering News on the Internet

Melika Yaghoubi

Supervisor: Dr. Hamid Hassanpour

Adviser: Dr. Morteza Zahedi

January ۲۰۱۳