سال پنجم، شماره‌های ۱۷ و ۱۸، پاییز و زمستان ۱۳۹۲
صص: ۹۶- ۷۳

فصلنامه علمی- پژوهشی
فناوری اطلاعات و ارتباطات ایران

خوشه‌بندی اسناد، مبتنی بر آنتولوژی و رویکرد فازی

مریم امیری* حسن ختن‌لو*
*کارشناس ارشد، دانشگاه بوعلی‌سینا، گروه کامپیوتر، همدان
**هیئت علمی، دانشگاه بوعلی‌سینا، گروه کامپیوتر، همدان
تاریخ دریافت: ۱۳۹۲/۰۱/۱۵ تاریخ پذیرش:۱۳۹۲/۰۳/۲۲

چکیده

داده‌کاوی، شناسایی و پردازش اطلاعات مفید از اسناد است که اساس آن بر مدل نمایش مفهومی اسناد، محاسبهٔ شباهت بین اسناد و استفاده از آن‌ها در خوشه‌بندی و دسته‌بندی اسناد، بازیابی و استخراج اطلاعات استوار است. در این مقاله روش نوینی برای نمایش آنتولوژیکال و مفهومی اسناد به صورت سلسله مراتبی ارائه شده است. با توجّه به آنتولوژی دامنهٔ مورد نظر، گراف مفهومی از سند ایجاد می‌شود. بر اساس این گراف آنتولوژیکال معیار شباهت متناسبی نیز ارائه شده است که فاصله و شباهت بین اسناد را بر اساس این نوع نمایش مشخص می‌نماید. در گام سوم سیستم استنتاج فازی با سه ورودی و یک خروجی طراحی شده است. این سیستم بر اساس سه شباهت ورودی، مقدار شباهت نهایی را تخمین می‌زند. در نهایت بر اساس ماتریس شباهت اسناد، الگوریتم خوشه‌بندی سلسله مراتبی پایین به بالا به منظور خوشه‌بندی اسناد اعمال می‌شود. برای ارزیابی الگوریتم پیشنهادی، نتایج با نتایج حاصل از روش‌های naïve Bayes ، دو الگوریتم مبتنی بر آنتولوژی و یک الگوریتم آماری مقایسه شده است. نتایج به دست آمده نشان می‌دهند که روش پیشنهاد شده مقادیر F-measure و Accuracy را بهبود می‌دهد. همچنین مقادیر FP و Error به میزان قابل توجّهی کاهش می‌یابد.

واژه‌های کلیدی: گراف مفهومی اسناد، ساختار آنتولوژیکال، آنتولوژی، معیار شباهت، ساختار سلسله مراتبی.

مقدمه

با رشد روز افزون اسناد روی وب، نیاز به مدیریت اسناد نیز بیش‌تر می‌شود. از نتایج رشد بیش از حدّ اسناد، مشکل بازیابی بیش از حدّ اطلاعات است. حل مشکل بازیابی بیش از حدّ اطلاعات شامل پردازش‌هایی نظیر جمع‌آوری اطلاعات، فیلتر کردن اطلاعات، بازیابی اطلاعات، استخراج اطلاعات، خلاصه‌سازی، خوشه‌بندی و دسته‌بندی اسناد است. هدف این پردازش‌ها کمک به کاربران برای یافتن اسناد مورد نیاز آن‌ها است. این پردازش‌ها وظایف اساسی را در زمینهٔ داده‌کاوی ایجاد می‌نمایند [۱]. داده‌کاوی، استخراج اطلاعات ضمنی، ناشناخته و مفید، تعریف شده است [۲].

داده‌کاوی، آنالیز داده از جنبه‌های مختلف و خلاصه‌سازی آن‌ها به صورت اطلاعات مفید تعریف شده است.

کاوش اسناد، شناسایی اطلاعات ناشناخته و استخراج آن‌ها از متون است [۳]. کاربردهای متعددی در زمینهٔ بازیابی اطلاعات وجود دارد که یکی از این کاربردها دسته‌بندی اسناد است. هدف از خوشه‌بندی[1] تقسیم یک مجموعهٔ بدون ساختار از اشیاء به داخل خوشه‌ها است، به گونه‌ای که اشیاء داخل خوشه تا جای ممکن به یدیگر مشابه باشند و از اشیاء داخل خوشه‌های دیگر متفاوت باشند.

1. Clustering

نویسندهٔ عهده‌دار مکاتبات: مریم امیری maryam_amiri2005_1365@yahoo.com

خوشه‌بندی اسناد یک رویکرد مهم برای سازمان‌دهی بدون سرپرست اسناد، استخراج خودکار موضوعات و بازیابی سریع اطلاعات است. برای مثال، موتورهای جستجوی وب هزاران صفحه را در پاسخ به یک پرس‌وجو بر می‌گردانند و کاربر را برای یافتن اطلاعات مرتبط دچار مشکل می‌سازند. خوشه‌بندی اسناد می‌تواند برای گروه‌بندی خودکار اسناد بازیابی شده به گروه‌های با معنی استفاده شود. به طور مشابه دسته‌بندی می تواند از قبل صورت گیرد و پردازش پرس وجو را آسان‌تر نماید بدین صورت که تنها نزدیک‌ترین خوشه‌ها به پرس‌وجو، جستجو می شوند [۴]. مدیریت مجموعه‌ای بزرگ از اسناد به تعدادی از خوشه‌ها، تأثیر و کارایی کاربردهای مبتنی بر متون که به سرعت و کیفیت بالایی نیاز دارند را بهبود می‌بخشد و مکمل خوبی برای موتورهای جستجو که اسناد بسیاری را برمی‌گردانند است [۵].

چارچوب این مقاله روشی جدید در تولید یک گراف مفهومی از اسناد بر اساس آنتولوژی دامنهٔ مورد نظر است. بر اساس این نمایش یک معیار اندازه‌گیری شباهت جدید تعریف شده است تا سطوح مشترک و متفاوت اسناد به طور دقیق‌تری شناسایی شوند تا در نهایت بتوان دقت روال‌های کاوش اسناد مبتنی بر مفهوم و آنتولوژی را بهبود داد.

در ادامهٔ مقاله در بخش ۲، مروری بر مدل‌های نمایش اسناد و معیارهای شباهت و در بخش ۳ به نمایش نوین آنتولوژیکال پیشنهادی پرداخته شده است. بخش ۴ به معیار شباهت متناسب با نمایش آنتولوژیکال پرداخته است. سیستم استنتاج فازی در بخش ۵ مطرح شده است. ارزیابی روش پیشنهادی، کارهای آتی و نتیجه‌گیری به ترتیب در بخش های ۶ و ۷ بیان گردیده‌اند.

- **مروری بر مدل‌های نمایش اسناد و معیارهای شباهت**

روش‌های بازیابی اطلاعات به دو دستهٔ عمده تقسیم می‌شوند: روش‌های آماری و روش‌های معنایی. در روش‌های معنایی تا حدّی آنالیز معنایی و نحوی صورت می‌پذیرد. به عبارت دیگر، سعی بر این است که متون زبان طبیعی که کاربر فراهم کرده است تا حدّی فهمیده شود. در روش‌های آماری اسنادی که بازیابی می‌شوند یا رتبه‌بندی بالایی دارند، اسنادی هستند که از لحاظ اندازه‌گیری آماری بررسی می‌شوند.

در روش‌های آماری فرض می‌شود کلمات می‌توانند نمایش معقولی از محتوای اسناد ارائه دهند. هیچ اطلاعی از ترتیب کلمات وجود ندارد. بنابراین به این روش‌ها مدل‌های bag of words گفته می‌شود. البته محتوای واقعی یک سند چیزی غیر از کلمات ظاهر شده است. روش‌های آماری به چندین دسته تقسیم می‌شوند: دودویی[۲]، دودویی بسط یافته[۳]، فضای بردار[۴] و احتمالاتی[۵]. روش‌های آماری اسناد را به چندین کلمه[۶] تجزیه می‌کنند. کلمات جمعیتی هستند که شمرده می‌شوند و به صورت آماری اندازه‌گیری می‌شوند. کلمات غالباً متحمّل پیش پردازش می‌شوند. آن‌ها معمولاً برای استخراج ریشه، ریشه‌یابی می‌شوند [۶] که هدف آن حذف تغییراتی است که به دلیل رخداد حالات مختلف دستوری در یک کلمه ایجاد می‌شود. روش دیگر پیش پردازش، حذف کلمات مشترک است که توانایی کمی برای جداسازی اسناد مرتبط و غیر مرتبط دارد. بنابراین موتورهای جستجو لیستی از کلمات توقف[۷] یا نویز تهیّه می‌نمایند. هر دو پیش‌پردازش اشاره شده وابسته به زبان هستند. معمولاً اوزان عددی به کلمات موجود در اسناد و پرس‌وجو نسبت داده می‌شود. اوزان به یک کلمه مشخص در هر سند تخصیص می‌یابد. اوزان تخصیص یافته به کلمات، میزان اهمّیت آن نشانه[۸] را برای محاسبهٔ شباهت بین اسناد مشخص می‌نماید. بنابراین کلمات یکسان در اسناد مختلف اوزان مختلفی دریافت می‌کنند.

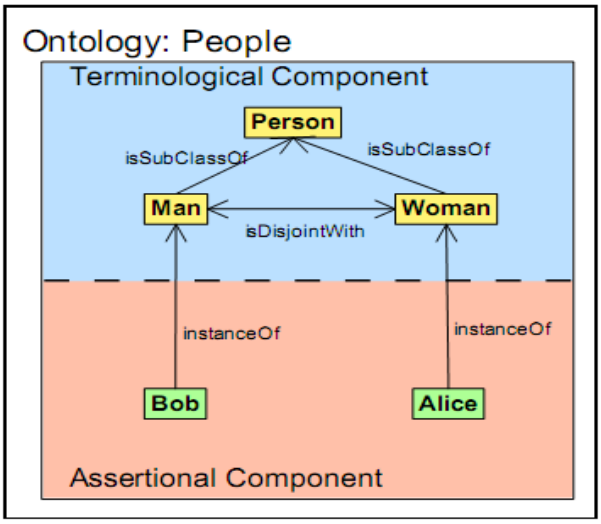
یک روش معمول نمایش اسناد و اندیس‌گذاری برای روش‌های آماری، نمایش اسناد متنی به صورت مجموعه‌ای از کلمات (عبارات یا n-grams) است. معمولاً کلمات از

2 .Boolean
3 .Extended Boolean
4 .Vector Space
5 .Probabilistic
6 .Term
7 .Stop Word
8 .Token

اسناد استخراج می‌شوند و در نهایت مجموعه‌ای از کلمات در دسترس می‌باشند که کل مجموعهٔ اسناد را نمایش می‌دهند. این مجموعه فضایی را تعریف می‌کند که هر کلمهٔ مجزا، یک بعد محسوب می‌شود [۷]. به هرکدام از کلمات موجود در سند، وزنی اختصاص می‌یابد که اهمّیت آن کلمه را برای جدا کنندگی سند مشخص می‌نماید. معمولاً در این حالت محتوای سند غیر از کلمات داخل بردار است. کارایی و سادگی این روش باعث شده است که اکثر موتورهای جستجو از این روش استفاده نمایند. واسط کاربری موتورهای جستجو، اغلب کلمات جستجو را در اسناد بازیابی شده با رنگ روشن نمایش می‌دهد که نشان دهندهٔ روش سادهٔ تناظر کلمات است.

روش دیگر نمایش اسناد، نمایش آن‌ها به صورت مجموعه‌ای از نشانه‌ها است. پایه‌ترین روش استفاده شده برای نمایش منابع متنی، مدل فضای بردار[۹] (VSM) است. در این مدل هر سند با یک بردار مشخص می‌شود. هر درایه از بردار منعکس کنندهٔ یک مفهوم خاص است. مقدار هر عنصر اهمّیت آن نشانه در نمایش معنایی سند است. پایگاه داده‌ای شامل d سند است که با t ترم[۱۰] توصیف شده‌اند، بنابراین به صورت یک ماتریس $d \times t$ نمایش داده می‌شود. هر سطر از ماتریس متناظر با بردار اسناد است. بدین گونه عناصر ماتریس، a_{ij} فرکانس وزن‌داری است که نشانه j در سند i اتفاق می‌افتد. در حالت دیگر از VSM، ستون‌های ماتریس بردار اسناد هستند و سطرهای ماتریس نشانه‌ها می‌باشند. مضمون معنایی پایگاه داده در فضای ستون‌های ماتریس گنجانده شده است به این معنا که بردارهای اسناد، آن مضمون را ایجاد کرده‌اند. در این نوع نمایش هر نشانه یک بعد مستقل است و در این فضا می‌توان هر سند را به صورت نقطه‌ای نمایش داد [۸]. شباهت‌ها یا تفاوت‌های بین اسناد را می‌توان فاصلهٔ بین نقاط تعریف نمود. در این نوع نمایش، شباهت‌ها بر اساس ظهور ترم‌های مشترک اندازه‌گیری می‌شوند یعنی اسنادی که ترم‌های مشترک بیش‌تری دارند شبیه‌تر از اسنادی هستند که ترم‌های مشترک کم‌تری دارند.اگرچه غالباً نمایش مبتنی بر مجموعهٔ کلمات برای خوشه‌بندی اسناد استفاده می‌شود، این روش نامناسب است به این دلیل که ارتباط بین کلماتی که با هم تکرار نمی‌شوند را نادیده می‌گیرد. بدیهی است که جملات بامعنی از کلمات بامعنی تشکیل شده است و هر سیستمی که بخواهد کار پردازش زبان طبیعی (NLP) را شبیه انسان انجام دهد باید دربارهٔ کلمات و معانی آن‌ها اطلاعات داشته باشد. این اطلاعات معمولاً از طریق فرهنگ لغت‌ها به خصوص WordNet [۹] فراهم می‌شود. نمایش مرسوم متون، مبتنی بر کلماتی هستند که در اسناد اتفاق افتاده‌اند و روش‌های دسته‌بندی، شباهت بین بردارها را محاسبه می‌نماید. ممکن است تعدادی از اسناد با اینکه کلمهٔ مشترکی ندارند، شامل اطلاعات معنایی مشابه باشند. برای حلّ چنین مشکلی استفاده از مدل مبتنی بر مفاهیم [۱۰] با استفاده از آنتولوژی ضروری است. روش‌های زیادی برای تعریف آنتولوژی وجود دارد. در [۱۱]، آنتولوژی یک چارچوب مفهومی برای تعریف کلاس‌های پایه از موجودیت‌های دامنهٔ دانش می‌باشد، رابطهٔ این نمونه‌ها با یکدیگر و مدیریت مفاهیم برحسب مفاهیم سطوح بالاتر همان طبقه‌بندی در طبیعت است. اصطلاح آنتولوژی برای ارجاع به محدّوده‌ای از منابع مفهومی و زبانی برای طبقه‌بندی و نمایش رسمی دانش که ممکن است استنتاج اتوماتیک و یا انواع خاصی از استدلال را پشتیبانی کند، استفاده شده است. مطابق با تعریف، آنتولوژی مجموعه‌ای از مفاهیم و روابط بین آن‌ها است که یک دیدگاه خلاصه از دامنه موضوع را فراهم می‌نماید [۱۱]. استفاده از آنتولوژی در داده‌کاوی برای خوشه‌بندی و دسته‌بندی اسناد و یادگیری الکترونیک است. شکل۱ آنتولوژی مردم را نشان می‌دهد که شامل مفاهیم، نمونه‌ها، روابط مابین آن‌ها است. در [۱۲] یک نمایش مفهومی مبتنی بر آنتولوژی ارائه شده است که معنی هر متن را به یک گراف بدون حلقهٔ مستقیم نگاشت می‌کند. [۱۳] پس از ساختن آنتولوژی دامنهٔ مورد نظر، سیستم با مجموعه‌ای از اسناد آموزش داده می‌شود. جملات برای استخراج POS (part of speech) و Chunk برچسب‌گذاری می‌شوند. پس از آن کارشناسان دامنهٔ کلمات را به مفاهیم آنتولوژی نگاشت می‌کنند. به کمک این مجموعهٔ آموزشی می‌توان نمایش مفهومی مجموعهٔ اسناد تست را نیز ساخت.

9 .Vector Space Model
10 .Term



شکل ۱- نمایش آنتولوژی شامل مفاهیم، نمونه‌ها و روابط [۱۴]

روش‌های مهمی برای محاسبهٔ شباهت بین دو سند d_i, d_j پیشنهاد شده است. یکی از متداول‌ترین روش‌ها ضرب داخلی دو بردار است. ضرب داخلی تابع کوسینوسی است که در رابطهٔ ۱ نشان داده شده است

$$\cos(d_i, d_j) = d_i^t d_j \Big/ \|d_i\| \|d_j\| \qquad (۱)$$

ضرب داخلی بیان کنندهٔ زاویهٔ بین آن‌ها است. اگر حاصل یک باشد، زاویه صفر درجه است و بیش‌ترین تطابق در این حالت وجود دارد. اگر حاصل صفر شود زاویهٔ بین آن‌ها نود درجه است و کم‌ترین تطابق وجود دارد. این اندازه‌گیری مشکلات خاصی برای اسناد با طول بلند ایجاد می‌کند.

به جز ضرب داخلی یا شباهت کسینوسی، توابع دیگری نیز برای اندازه‌گیری شباهت وجود دارد، خانواده‌ای از فاصله‌ها [۱۵] به صورت رابطهٔ ۲ بیان شده‌اند. این تابع، فاصله را بر حسب اجزای دو سند محاسبه می‌نماید. جدا از این معیارهای فاصله، توابعی هستند که تنها تعداد کلمات مشترک و تعداد کلمات غیر مشترک را شمارش می‌کنند. یکی از این توابع معروف ضریب Dice است [۱۶] که در رابطهٔ ۳ نشان داده شده است:

$$L_p(D_1, D_2) = [\sum_i |d_{1i} - d_{2i}|^p]^{1/p} \qquad (۲)$$

$$Dice = \frac{2w}{n_1 + n_2} \qquad (۳)$$

که مقدار w تعداد کلمات مشترک بردارهای اسناد است، n_1 تعداد کلمات با فرکانس غیر صفر در اولین سند و n_2 تعداد کلمات با فرکانس غیر صفر در دومین سند است. مخرج کسر نوعی نرمال‌سازی است. یک تابع معمول دیگر برای اندازه‌گیری شباهت، ضریب Jaccard است [۱۶]. این تابع شباهت در رابطهٔ ۴ ذکر شده است:

$$Jaccard(D_1, D_2) = \frac{w}{N - z} \qquad (۴)$$

$$Jaccard(D_1, D_2) = \frac{w}{N - z} \qquad (۴)$$

که w تعداد کلمات مشترک اسناد D_1 و D_2 است و N تعداد کل کلمات مجزا در فضای برداری و z تعداد کلمات مجزایی هستند که نه در D_1 و نه در D_2 است. بنابراین مخرج کسر تعداد کلماتی را نشان می‌دهند که در D_1 یا در D_2 و یا در هر دو رخ می‌دهند. توابع محاسباتی بالا برای محاسبهٔ شباهت اسناد، فرض می‌کنند که مجموعهٔ اسناد ایستا هستند. در حالت "routing"، اسناد دارای یک

جریان ورودی متغیر هستند و هر سند باید برای یکی از N گروه متناظر با N موضوع از پیش تعریف شده، "routed" (مسیریابی)، شود.

محاسبهٔ شباهت‌ها در خوشه‌بندی به طور مکرر انجام می‌گیرد. بنابراین با سریع انجام دادن آن می‌توان به کل روال سرعت بخشید. تعیین شباهت‌ها نیز بستگی به تعداد کلمات به کار رفته دارد، پس با کاهش تعداد کلمات در نمایش می‌توان سرعت را افزایش داد [۱۷]. روش‌های پیچیده‌ای برای کاهش تعداد کلمات وجود دارد. در [۱۸] روش‌های انتخاب نشانه، مبتنی بر تغییرات فرکانس نشانه، در متن پیشنهاد شده است. در [۱۹] روال کاهش نشانه‌ها افکنش[11] نامیده می‌شود که فضای برداری به یک فضای جدید با تعداد ابعاد کم‌تر تصویر می‌شود. همچنین داده‌های متنی مشکل بزرگی دارند: مشکل ابعاد زیاد. در فضاهایی با ابعاد بالا، فاصلهٔ بین هر جفت از نقاط تقریباً برای انواع گوناگون توزیع داده‌ها و توابع فاصله یکسان است [۲۰] که این انگیزه‌ای برای کاهش ابعاد دادهٔ ورودی است. تعدادی از روش‌های انتخاب ویژگی، به جستجوی زیر مجموعه‌های ویژگی و ارزیابی هر کدام از این مجموعه‌ها با استفاده از معیارهایی پرداخته‌اند [۲۱]. همبستگی در میان ابعاد، اغلب مخصوص محل و موقعیت داده است، به این معنی که تعدادی از نقاط داده با یک مجموعه از ویژگی‌ها و سایر آن‌ها با ویژگی‌های متفاوت وابسته هستند. یعنی در فضایی با ابعاد بالا شبیه داده‌های متنی، هر خوشه ساختار زیر فضایی خاص خودش را دارا است.

یادگیری با هر دو دادهٔ برچسب خورده و بدون برچسب، یادگیری نیمه سرپرست نامیده می‌شود. این روش اخیراً با توجّه زیادی مورد مطالعه قرار گرفته است و اساساً به صورت یک روش استخراج اطلاعات در داده‌های بدون برچسب، برای بهبود کارایی مدل خوشه‌بندی استفاده شده است [۵]. در نمایش گرافی و آنتولوژیکال اسناد روشی که برای اندازه‌گیری شباهت به صورت معمول استفاده شده است بر اساس‌عبارت ساده‌ی $X = \frac{ST}{TT}$ است که در آن ST تعداد مفاهیم مشترک و TT مجموع کل مفاهیم دو سند است [۱۳].

- **نمایش نوین آنتولوژیکال**

روش پیشنهادی به تولید یک گراف وزن‌دار آنتولوژیکال می‌پردازد. با توجّه به مضمون و مفهوم اسناد، مفاهیم اصلی شناسایی می‌شوند و با توجّه به اهمّیت‌شان در سند اوزانی دریافت می‌کنند. سپس ساختار مفهومی سند شناسایی می‌شود و مفاهیم شناسایی شده در مرحلهٔ قبل با توجّه به این ساختار با یال‌های جهت‌دار و وزن‌دار به یکدیگر متصل می‌گردند. در ادامه روش پیشنهادی با جزئیات بیش‌تری مطرح می‌شود.

۱. پیش پردازش اولیه

یک پاراگراف مجموعه‌ای از چند جمله است که راجع به یک مفهوم خاص بحث می‌نماید. اساس روش پیشنهادی با توجّه به این نکته است. در مرحلهٔ پیش پردازش، پاراگراف‌ها واحدّهای پردازشی هستند. ابتدا متن به پاراگراف‌هایش تجزیه می‌شود. سپس برای هر پاراگراف عملیات پیش‌پردازشی نظیر نشانه‌گذاری[12]، حذف کلمات نویزی و ریشه‌یابی صورت می‌گیرد. در نهایت برای هر پاراگراف دو مجموعه از نشانه‌ها نگهداری می‌شود: مجموعهٔ نشانه‌های اصلی و مجموعهٔ ریشه‌یابی شدهٔ نشانه‌های اصلی.

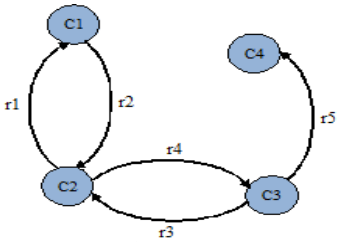
۲. نگاشت کلمات به مفاهیم آنتولوژی

مسئله‌ای که همیشه به عنوان یک چالش مطرح بوده است چگونگی استخراج اطلاعات از آنتولوژی است. یکی از روش‌هایی که می‌توان از آن استفاده کرد این است که به طور مستقیم با استفاده از یک زبان پرس‌وجوی آنتولوژی، اطلاعات را از آنتولوژی استخراج نمود. روش دوم تبدیل آنتولوژی به نوعی پایگاه داده‌ای مانند RDBMS است تا بتوان با استفاده از زبان‌های پرس‌وجو، اطلاعات را از این انباره بیرون کشید. در الگوریتم پیشنهادی از روش دوم یعنی نگاشت آنتولوژی به پایگاه داده برای استخراج اطلاعات استفاده شده است. به این ترتیب که آنتولوژی از یک فایل

11 .Projection

12 .Tokenization

OWLاستخراج می‌شود و سپس در یک پایگاه داده رابطه‌ای ذخیره می‌شود.جدول ایجاد شده به چندین جدول کوچک‌تر شامل کلاس‌ها، نمونه‌ها، ماتریس کلاس- کلاس و ماتریس کلاس- نمونه تبدیل می‌شود. جدول کلاس، شامل مفاهیم موجود در گراف آنتولوژی است. جدول نمونه‌ها، تمامی نمونه‌های موجود در آنتولوژی را در بر می‌گیرد. شکل ۲ نمونه‌ای از یک گراف آنتولوژی را با چهار مفهوم نمایش می‌دهد که در آن گره‌های c_1، c_2، c_3، c_4 نمایانگر مفاهیم آنتولوژی هستند. ارتباط بین کلاس‌ها با r_iها نمایش داده شده است. برای تجزیه و تحلیل آنتولوژی می‌توان گراف آنتولوژیکال را توسط یک ماتریس مجاورت[13] با سطر و ستون مساوی نمایش داد که در آن سطرها و ستون‌ها نمایانگر نودهای گراف یا همان کلاس‌های آنتولوژی هستند. اعداد قرار گرفته در هر درایه ماتریس نیز نشان دهنده تعداد ارتباطات بین کلاس‌های مربوطه خواهند بود [۲۲].با استفاده از دیکشنری معکوس[14] جدول دیگری از مفاهیم – کلمات ساخته شده است. به ازای مفاهیم موجود در آنتولوژی، صد کلمهٔ مرتبط با هر کدام از این مفاهیم از دیکشنری یافت می‌شود و در پایگاه داده ذخیره می‌شوند. کلمات استخراج شده از دیکشنری مجدداً توسط کارشناسان دامنه بررسی می‌شوند و مرتبط‌ترین کلمات برای هر مفهوم انتخاب می‌شوند.برای نگاشت کلمات به مفاهیم، با توجّه به مفاهیم موجود در آنتولوژی، تا چندین سطح از مفاهیم بررسی می‌شوند. در ابتدا مفاهیم مستقیم بررسی می‌شوند. منظور از مفاهیم مستقیم مفاهیمی هستند که عبارت (نشانه) موجود در سند، مفهومی در آنتولوژی است. به ازای مفاهیم مستقیم تا چند سطح از فرزندان و پدران مفهوم یافت شده نیز در نظر گرفته می‌شوند. به این مفاهیم، مفاهیم غیرمستقیم نوع ۱ گفته می‌شود.



شکل ۲- نمونه‌ای از یک گراف آنتولوژی

اگر عبارت جزء مفاهیم مستقیم نبود به عنوان یک نمونه بررسی می‌شود که در این صورت تا چند سطح از پدران این نمونه نیز به مجموعهٔ مفاهیم غیرمستقیم-۱ (مفاهیم غیرمستقیم نوع ۱) اضافه می‌شوند. در صورتی که عبارت مورد نظر مفهوم مستقیم یا نمونه نبود مفاهیم غیرمستقیم نوع ۲ (مفاهیم غیرمستقیم-۲) بررسی می‌شوند. این نوع مفاهیم غیرمستقیم از جستجو در جدول مفاهیم-کلمات حاصل می‌شوند. برای این مفاهیم نیز پدران و فرزندان تا چند سطح به مفاهیم غیرمستقیم-۱ اضافه می‌شوند. در نهایت برای هر پاراگراف مجموعه‌ای از مفاهیم مستقیم و غیرمستقیم نوع ۱و۲ با تعداد دفعات ارجاع به آن‌ها موجود است. برای مفاهیم غیرمستقیم-۱ فاصله تا مفهوم مستقیم و همچنین تعداد دفعاتی که به عنوان والد یا فرزند انتخاب شده است نیز در نظر گرفته می‌شود.

۳. رفع ابهام از مفاهیم

در بخش قبل به مفاهیم غیرمستقیم نوع ۱و۲ اشاره شد. مشکلی که ممکن است در رابطه با این مفاهیم پیش آید ابهام است. یک مفهوم مستقیم ممکن است چندین والد یا فرزند به عنوان مفهوم غیرمستقیم نوع ۱ داشته باشد. همچنین یک کلمه ممکن است متناظر با چندین مفهوم باشد. در کلیهٔ این حالات مشکل ابهام پیش می‌آید. برای این مفاهیم تعداد دفعاتی که به صورت مبهم شناسایی شده‌اند نگه‌داری می‌شود. پس از استخراج تمام مفاهیم یک پاراگراف، باید مفاهیم مبهم بررسی شوند و مناسب‌ترین آن‌ها باقی بمانند. روشی که برای رفع ابهام در این الگوریتم پیشنهاد شده است بدین ترتیب است که با توجّه به سایر مفاهیم مستقیم و غیرمستقیم غیرمبهم هر پاراگراف، مفاهیم مبهم پاراگراف رفع ابهام می‌شوند و مناسب‌ترین مفاهیم انتخاب می‌شوند. روش رفع ابهام بدین ترتیب است که ابتدا اهمّیت مفهوم مبهم در پاراگراف مزبور مشخص می‌شود. اگر تعداد ارجاعات بدون ابهام به یک مفهوم از ۰/۷ کل تعداد ارجاعات کم‌تر باشد، آنگاه از این مفهوم باید رفع ابهام کرد. در رابطه ۵، m اهمّیت مفهوم مبهم را مشخص می‌کند. اگر OC مفهومی باشد که برچسب مبهم داشته باشد، آنگاه $oc.counter$ کل تعداد دفعات ارجاع به این مفهوم، $oc.tag$

13 .Adjacency Matrix

14 .Inverse Dictionary

تعداد دفعاتی که این مفهوم به صورت مبهم شناسایی شده و oc.distance فاصلهٔ این مفهوم غیرمستقیم مبهم تا مفاهیم مستقیم اصلی است.

$$m = 0.7 \times (oc.counter - oc.tag) / |oc.distance|$$

اگر مفهومی کاملا مبهم باشد یعنی $m = 0$، به آن مقدار حداقلی غیر از صفر نسبت داده می‌شود. سپس این مفهوم نسبت به سایر مفاهیم مستقیم و غیرمستقیم غیرمبهم ارزیابی می‌شود. رابطهٔ ۶ ارزیابی نسبت به مفاهیم مستقیم را نشان می‌دهد که در آن منظور از $OC2$ مفهوم مستقیم و OC مفهوم غیرمستقیم مبهم است. تابع find_distance فاصلهٔ بین دو مفهوم در آنتولوژی را می‌یابد و تابع rout_number تعداد مسیرهای ممکن بین دو مفهوم را مشخص می‌نماید. رابطهٔ ۶ برای تمامی مفاهیم مستقیم محاسبه می‌شود و نتایج با یکدیگر جمع می‌شوند.

$$mark = oc2.counter \times (2 / find\_distance(oc, oc2)) \times rout\_number(oc, oc2) \times m$$

رابطهٔ ۷ ارزیابی مفاهیم مبهم نسبت به مفاهیم غیرمستقیم غیرمبهم را نشان می‌دهد. که $OC2$ مفهوم غیرمستقیم غیرمبهم و OC مفهوم غیرمستقیم مبهم است. این رابطه برای تمامی مفاهیم غیرمستقیم غیرمبهم محاسبه می‌شود و نتایج با یکدیگر جمع می‌شوند.

نتایج حاصل از رابطهٔ ۶ و ۷ با یکدیگر جمع می‌شوند، اگر مقدار به دست آمده از یک آستانه بیش‌تر بود از این مفهوم رفع ابهام می‌شود در غیر این صورت مفهوم مورد نظر حذف می‌شود. اگر نسبت مفاهیم رفع ابهام شده به کل مفاهیم مبهم پاراگراف از یک آستانه مشخص کم‌تر باشد مفاهیم غیرمبهم یک پاراگراف قبل و یک پاراگراف بعد نیز بررسی می‌شوند و راجع به مفاهیم مبهم پاراگراف فعلی تصمیم گرفته می‌شود. این امر سبب می‌شود، اگر پاراگراف فعلی از لحاظ مفاهیم غیرمبهم ضعیف باشد مفاهیم بی‌سبب حذف نشوند و با دقت بیش‌تری بررسی شوند.

$$mark = oc2.counter \times 1 / |oc2.distance| \times (1.5 / find\_distance(oc, oc2)) \times rout\_number(oc, oc2) \times m \quad (۷)$$

۴. استخراج ساختار سلسله مراتبی مفهومی اسناد

پس از انجام مراحل فوق در نهایت برای هر پاراگراف مجموعه‌ای از مفاهیم مستقیم و غیرمستقیم و اطلاعاتی راجع به آن‌ها در دسترس است. مرحلهٔ آخر وزن‌دهی به مفاهیم و ترسیم شمای گرافی سند است. به دلیل اهمّیت مفاهیم مستقیم و مفاهیم غیرمستقیم-۲ در مضمون سند این مفاهیم با ضرایب بیش‌تری نسبت به سایر مفاهیم در نظر گرفته می‌شوند. برای وزن‌دهی به مفاهیم سند، تعداد ارجاعات هر مفهوم به تعداد کل ارجاعات نوع مفهوم مورد نظر تقسیم می‌شود و برای هر مفهوم با توجّه به نوع آن، وزن منحصر به فردی در نظر گرفته می‌شود. پس از آن مفاهیم مستقیم و نمونه‌هایی که در سند بوده‌اند انتخاب می‌شوند و به کمک ماتریس‌های کلاس- کلاس و کلاس-نمونه، روابط بین آن‌ها استخراج می‌شود و ماتریس روابط نظیر هر سند ایجاد می‌شود. بنابراین نودهای گراف و اوزان مربوط به آن‌ها ایجاد می‌شود.

به منظور ترسیم یال‌های گراف و محاسبهٔ اوزان آن‌ها، کلیهٔ مفاهیم غیرمستقیم-۲ به صورت مفهوم مستقیم در نظر گرفته می‌شوند و مفاهیم غیرمستقیم-۱ که همنام با مفاهیم مستقیم هستند، بررسی می‌شوند. اگر این مفاهیم غیرمستقیم همنام به عنوان فرزند انتخاب شده باشند بنابراین جهت یال‌ها باید از مفاهیم مستقیم پدر این مفهوم به مفهوم مستقیم مورد نظر باشد. اگر مفاهیم غیرمستقیم به عنوان پدر انتخاب شده‌اند، جهت یال‌ها باید از فرزندان مستقیم این مفهوم به مفهوم مربوطه باشند. برای محاسبهٔ اوزان نظیر یال‌های ترسیم شده در گراف، از ماتریس روابط

نظیر سند استفاده می‌شود. اوزان مفاهیم غیرمستقیم همنام، در تعداد رابطهٔ بین دو مفهوم مورد نظر ضرب می‌شود و بر تعداد ارتباط گره مورد نظر با پدرانش (یا فرزندانش) تقسیم می‌شود. همچنین متناسب با فاصلهٔ مفاهیم همنام غیرمستقیم، ضریبی در نظر گرفته می‌شود. در رابطهٔ ۸ هدف بررسی پدران با فاصلهٔ یک است. اگر i مفهوم مستقیمی باشد که همنام با آن، مفهوم پدر غیرمستقیم با فاصلهٔ یک نیز وجود داشته باشد، برای تمامی فرزندان آن، j، یال‌های j به i به صورت زیر محاسبه می‌شود که sum_child_i بیانگر تعداد فرزندان مفهوم i با فاصلهٔ یک است، $concept_i.weight$ وزن مفهوم غیرمستقیم همنام با مفهوم اصلی است و $matrix[i,j]$ بیانگر تعداد ارتباط بین j و i است. ضریب W را نیز می‌توان با توجّه به فاصلهٔ مفاهیم غیرمستقیم مقدار داد. رابطهٔ ۸ برای فرزندان نیز به طریق مشابه نوشته می‌شود.

$$weight_{j-i} = (concept_i.weight \times W \times matrix[i,j]) / sum\_child_i$$

در نهایت پس از محاسبهٔ حالات مختلف از فرزندان و پدران با فواصل متفاوت، گراف جهت داری تولید می‌شود که نودهای آن مفاهیم مستقیم هستند. اوزان این مفاهیم و یال‌های مربوطه نیز از طریق توضیحات بیان شده و رابطهٔ ۸ محاسبه شده‌اند. گراف ایجاد شده برای هر سند به صورت ماتریس در پایگاه داده ذخیره می‌شود تا در مراحل محاسبهٔ ماتریس شباهت بین اسناد و کاوش آن‌ها، به راحتی قابل بازیابی باشند.

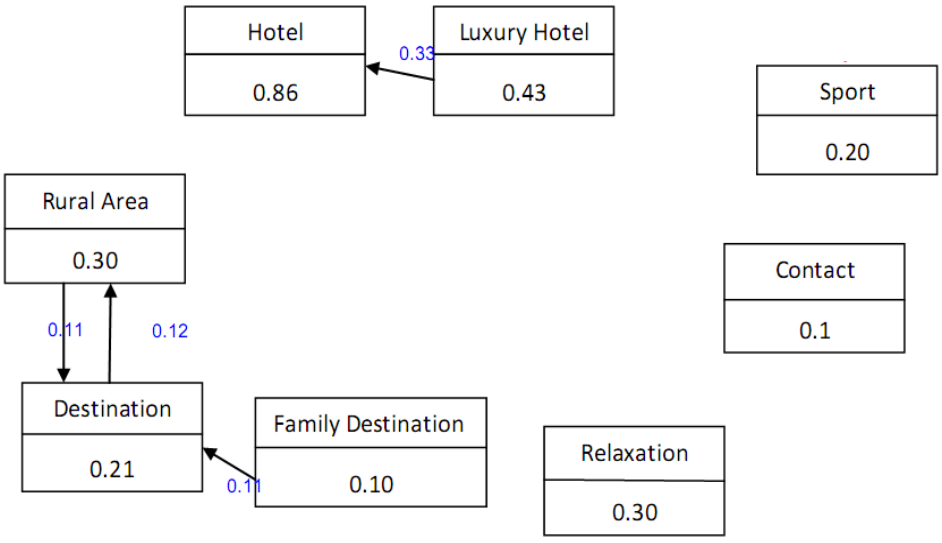
به عنوان یک مثال، شکل ۳ یک سند دلخواه در زمینهٔ هتل است. از آنتولوژی دامنه مورد نظر جداول و ماتریس‌های مذکور ساخته شده‌اند و طبق روال توضیح داده شده، شکل ۴ گراف آنتولوژیکال تولید شده را نمایش می‌دهد که شامل مفاهیم، اوزان و یال‌های آن‌ها است. با توجّه به اوزان مفاهیم، کلیت سند راجع به مفاهیم hotel و luxury hotel است. با توجّه به این اوزان می‌توان به سادگی مفاهیم کلی و پراهمّیت و مفاهیم جزئی را مشخص نمود. به عبارت دیگر این اوزان را می‌توان همانند درجهٔ عضویت فازی تفسیر نمود که سند با چه درجه عضویتی به هر مفهوم متعلق است. در مورد اوزان و جهات یال‌ها نیز می‌توان چنین تفسیری ارائه داد. در واقع با توجّه به آنتولوژی موجود، شمای گرافی سند، زیر مجموعه‌ای از همان آنتولوژی است که اوزان یال‌ها و نودها در آنتولوژی اصلی ۰.۱۰۰ است اما در اسناد با توجّه به مضمون و مفهوم سند این اوزان مقادیر مختلفی می‌گیرند.

Drake & Cavendish have an amazing selection of luxury hotels in Iran featured within this section that are perfect for a relaxing vacation to Iran. Many of the luxury hotels featured are also perfect for honeymoons, romantic breaks or special occasions. No matter what type of vacation you are planning we hope to have compiled a comprehensive list of the very best hotels in Iran for you to research. If you'd like to search for a hotels based on your specific requirements, we invite you to read our Specialist Collections section where we have grouped hotels by category rather than destination, this section is great if you know you want a boutique hotel, or a family friendly hotel. We'd like to think we've done the hard work for you by grouping them together!

شکل ۲- نمایش یک سند در زمینه هتل

- **معیار شباهت متناسب با نمایش آنتولوژیکال**

مهم‌ترین گام‌ها برای بهبود روال‌های کاوش اسناد، نمایش مفهومی مناسب و معیار شباهت متناسب با این نمایش است. هرچه معیار شباهت توانایی بیش‌تری برای تقریب سطوح اختلاف و تشابه بین اسناد داشته باشد، مناسب‌تر و کاربردی‌تر است. روش آنتولوژیکال پیشنهادی، دارای چهار جزء با معنی است که برای تعیین شباهت و تفاوت بین اسناد کلیدی می‌باشند: مفاهیم و اوزان هر مفهوم، یال‌ها و اوزان منتسب به هر یال. معیار پیشنهاد شده برای مفاهیم و یال‌ها به صورت مجزا محاسبه می‌شود و خروجی آن ماتریس‌های شباهت مجزا برای مفاهیم و یال‌ها است. در مراحل بعد مبتنی بر ماتریس شباهت محاسبه شده و با



شکل ۳- گراف آنتولوژیکال سند در شکل (۳)

استفاده از سیستم استنتاج فازی، قوانین فازی و یک الگوریتم خوشه‌بندی مناسب می‌توان نتایج کاوش اسناد را بهبود داد.

معیار پیشنهادی، درجهٔ عضویت، اولویت و اهمّیت هر مفهوم را در نظر می‌گیرد و براساس مفاهیم مشترک (یال‌های مشترک) هر دو سند میزان شباهت بین دو سند را تقریب می‌زند. برای هر مفهوم مشترک در هر دو سند، دو وزن w_1 و w_2 محاسبه می‌شود و در نهایت به کمک این اوزان میزان شباهت دو سند تقریب زده می‌شود. رابطه‌های ۹ و ۱۰ به ترتیب محاسبه اوزان w_1 و w_2 را بیان می‌نمایند که w_1 وزن مربوط به تفاوت اولویت و اهمّیت مفاهیم و w_2 مربوط به اختلاف اوزان مفاهیم مشترک دو سند است. در رابطهٔ ۹ $concept_i$ مفهوم مشترک $concept$ در سند i است، $order(concept_i)$ اولویت مفهوم $concept$ را در سند i مشخص می‌نماید، x و y دو سند دلخواه هستند که هدف محاسبه شباهت بین آن دو است و $\max\_length(x,y)$ حداکثر اختلاف اهمّیت مفاهیم مشترک بین دو سند تعریف شده است. در رابطهٔ ۱۰ $weight(concept_x)$ وزن مفهوم مشترک $concept$ در سند x است.

$$w_1 = (\max\_length(x,y) - \left|(order(concept_x) - order(concept_y))\right|) / \max\_length(x,y) \quad (۹)$$

$$w_2 = 1 - \left|weight(concept_x) - weight(concept_y)\right| \quad (۱۰)$$

رابطهٔ ۱۱ بیانگر معیار شباهت برای اندازه‌گیری شباهت بین دو سند x و y است. در این رابطه، m بیانگر تعداد مفاهیم مشترک دو سند است، نماد $||$ بیانگر اندازهٔ مجموعه (تعداد مفاهیم) بوده و $|x \cup y| = |x| + |y| - |x \cap y|$ است.

$$sim(x,y) = \frac{\sum_{i=1}^{m} w_1 w_2}{|x \cup y|} \quad (۱۱)$$

- **سیستم استنتاج فازی**

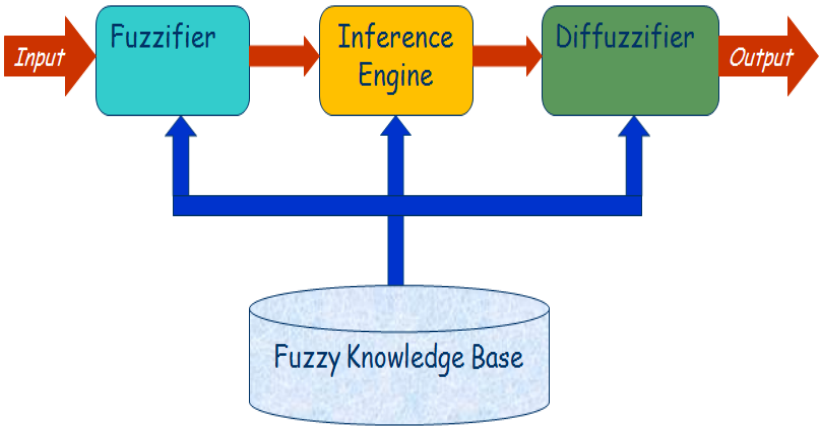
سیستم‌های فازی، سیستم‌های مبتنی بر دانش یا قواعد هستند؛ قلب یک سیستم فازی یک پایگاه دانش است که از قواعد اگر ـ آنگاه فازی تشکیل شده است. یک قاعده اگر ـ آنگاه فازی، یک عبارت اگر ـ آنگاه است که بعضی کلمات آن به وسیلهٔ توابع عضویت پیوسته مشخص شده‌اند. موتور استنتاج فازی، این قواعد را با یک نگاشت از مجموعه‌های فازی در فضای ورودی به مجموعه‌های فازی و در فضای خروجی بر اساس اصول منطق فازی ترکیب می‌کند.

سیستم استنتاج فازی متشکل از سه بخش فازی‌سازی، موتور استنتاج فازی و غیرفازی‌سازی است. در بخش فازی‌سازی یک متغیر crisp با استفاده از توابع عضویت تعریف شده به یک متغیر زبانی تبدیل می‌شود. در بخش دوم با استفاده از قوانین فازی (قوانین اگر-آنگاه) مقدار خروجی فازی تولید می‌شود. بخش غیرفازی‌سازی مقدار خروجی فازی از موتور استنتاج را با استفاده از توابع عضویت تعریف شده، به یک مقدار crisp تبدیل می‌نماید. این روند در شکل ۵ نشان داده شده است. سیستم استنتاج طراحی شده در این بخش دارای سه ورودی است: میزان شباهت مفاهیم کلی، میزان شباهت مفاهیم جزئی و میزان شباهت یال‌های موجود در شمای گرافی اسناد. مفاهیم جزئی و کلی هر سند به صورت نسبی مشخص می‌شود.

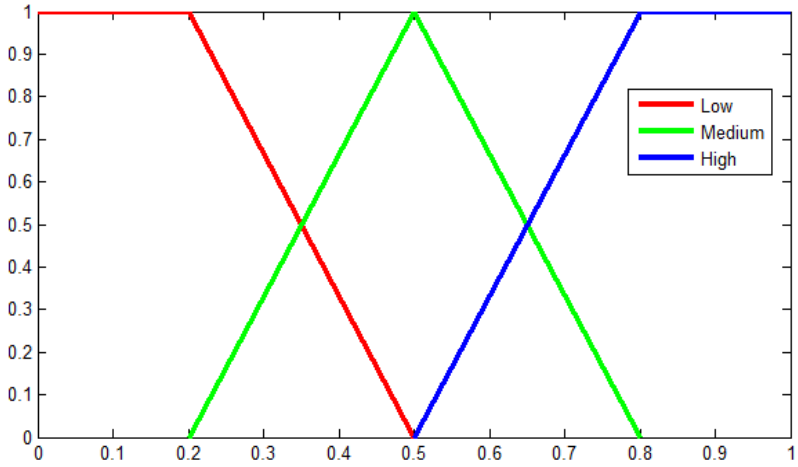
برای تعیین مفاهیم کلی و جزئی، ابتدا بیش‌ترین وزن موجود شناسایی می‌شود. سپس با استفاده از رابطهٔ (۱۲) مفاهیم جزئی و کلی برای هر سند مشخص می‌شوند که منظور از $\max$ مقدار وزن بیشینه در سند است و $co.weight$ مقدار وزن مفهوم co را مشخص می‌نماید.

برای هر یک از ورودی‌های سیستم استنتاج، سه تابع عضویت $High, Low, Medium$، مطابق با شکل ۶ مشابه با توابع عضویت در [۱۳]، تعریف شده است. محور افق بیان‌گر میزان شباهت بین اسناد و محور عمودی درجه عضویت را نشان می‌دهد. موتور استنتاج فازی ممدانی برای فازی‌سازی مقادیر ورودی استفاده شده است. مدل سیستم استنتاج فازی ممدانی از عملگر $\min-\min-\max$ استفاده می‌نماید.

$$\text{if } co.weight: \begin{cases} \geq 0.1, \geq \dfrac{\max}{2+(\max * 0.1)} : \text{main concept} \\ \geq 0.05, < \dfrac{\max}{2+(\max * 0.1)} : \text{detail concept} \end{cases} \qquad (۱۲)$$



شکل ۴- ساختار سیستم استنتاج فازی [۲۳]



شکل ۵- توابع عضویت ورودی فازی

موتور استنتاج طراحی شده به منظور خوشه‌بندی از بیست و هفت قانون فازی که در جدول ۱ بیان شده است، استفاده می‌نماید. هر سطر از این جدول بدین گونه تفسیر می‌شود (قانون۱):

if main_concept is high and detail_concept is high and main_edge is high then similarity1 is high

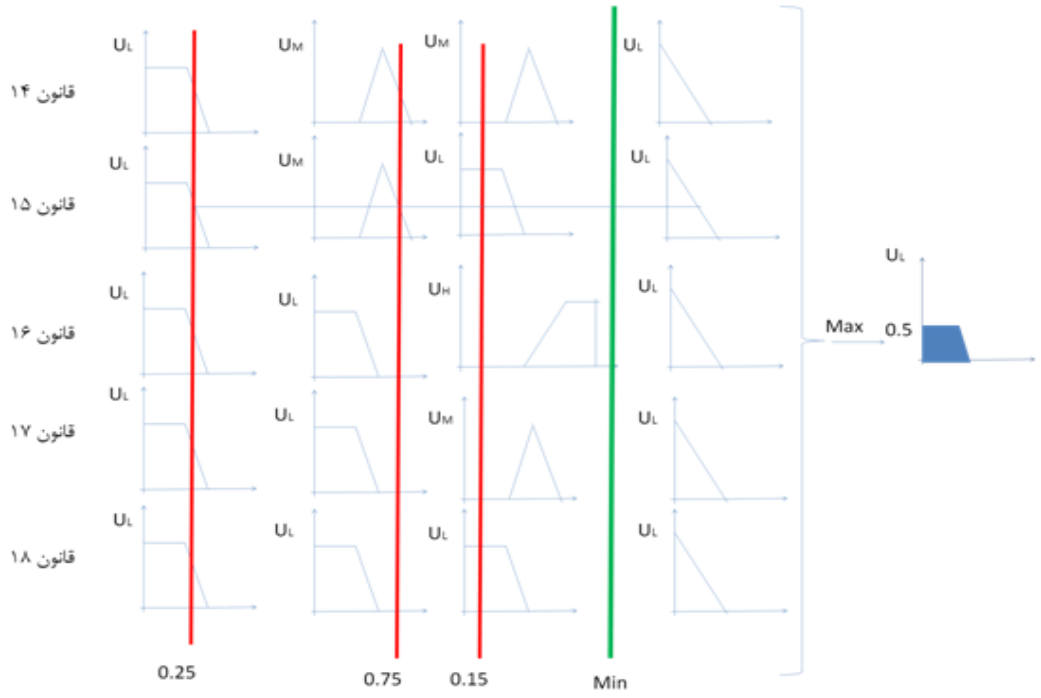
اگر خوشه‌بندی اسناد فقط در یک دامنهٔ مشخص باشد و هدف خوشه‌بندی دقیقی از اسناد یک دامنه باشد، ساختار مفهومی این اسناد در تعیین شباهت بین آن‌ها دارای اهمّیت است. اگر خوشه‌بندی در چند دامنه صورت گیرد و هدف یافتن یک خوشه‌بندی کلی بر اساس موضوعات دامنه باشد، ساختار مفهومی اسناد تأثیر چندانی در تعیین شباهت بین اسناد ندارند و در عوض مفاهیم کلی دارای اهمّیت بیش‌تری هستند. بنابراین اگر هدف خوشه‌بندی جزئی اسناد باشد، جدول similaity1 و اگر هدف خوشه‌بندی کلی باشد similaity2، به کار گرفته می‌شوند. خروجی سیستم فازی دارای سه تابع عضویت با مقادیر شباهت $High, Low, Medium$ است. در نهایت با استفاده از روش غیرفازی‌سازی مقدار شباهت نهایی بین دو سند تخمین زده می‌شود. بنابراین می‌توان مراحل زیر را برای محاسبهٔ شباهت بین اسناد بیان نمود:

محاسبهٔ شباهت بین مفاهیم جزئی، کلی و یال‌های اسناد استفاده از سیستم استنتاج فازی و تولید خروجی فازی غیرفازی‌سازی خروجی و محاسبهٔ مقدار نهایی شباهت دو سند خوشه‌بندی سلسله مراتبی اسناد بر اساس ماتریس شباهت نهایی برای غیرفازی کردن مقادیر شباهت خروجی دو مرحله وجود دارد. مرحلهٔ اول مشخص نمودن میزان شباهت بین اسناد (High,Low,Medium) است. دومین مرحله غیرفازی کردن مقدار این شباهت است. سیستم ممدانی برای مشخص نمودن میزان شباهت بین اسناد از روش Max گیری استفاده می‌نماید. در سیستم استنتاج فازی مطرح شده، پنج حالت ممکن است بین توابع عضویت متغیر خروجی رخ دهد که در رابطهٔ ۱۳ بیان شده‌اند. منظور از U_H میزان درجه عضویت به شباهت High، منظور از U_L میزان درجه عضویت به شباهت Low و منظور از U_M میزان درجه عضویت به شباهت Medium است.

جدول ۱- قوانین مربوط به سیستم استنتاج فازی

No	Main concept	Detailed concept	Main Edge	Similarity1	Similarity2
1	High	High	High	High	High
2	High	High	Medium	High	High
3	High	High	Low	High	High
4	High	Medium	High	Medim	High
5	High	Medium	Medium	High	High
6	High	Medium	Low	High	High
7	High	Low	High	Medium	High
8	High	Low	Medium	Medium	High
9	High	Low	Low	High	High
10	Low	High	High	Low	Medium
11	Low	High	Medium	Medium	Medium
12	Low	High	Low	Medium	Medium
13	Low	Medium	High	Low	Medium
14	Low	Medium	Medium	Low	Low
15	Low	Medium	Low	Medium	Low
16	Low	Low	High	Low	Low
17	Low	Low	Medium	Low	Low
18	Low	Low	Low	Low	Low
19	Medium	High	High	Medium	High
20	Medium	High	Medium	High	High
21	Medium	High	Low	High	High
22	Medium	Medium	High	Medium	High
23	Medium	Medium	Medium	Medium	High
24	Medium	Medium	Low	Medium	High
25	Medium	Low	High	Low	Medium
26	Medium	Low	Medium	Medium	Medium
27	Medium	Low	Low	Medium	Medium

$$S_{ij} = \begin{cases} \frac{2+\mathrm{U_M}}{6}, \text{if } \mathrm{U_L} > \mathrm{U_H}, \mathrm{U_M} > \mathrm{U_H}, \mathrm{U_M} > \mathrm{U_L} \\ \frac{4-\mathrm{U_M}}{6}, \text{if } \mathrm{U_L} < \mathrm{U_H}, \mathrm{U_M} > \mathrm{U_H}, \mathrm{U_M} > \mathrm{U_L} \\ \frac{1}{2}, \text{if } \mathrm{U_L} = \mathrm{U_H}, \mathrm{U_M} > \mathrm{U_H}, \mathrm{U_M} > \mathrm{U_L} \\ \frac{2+\mathrm{U_H}}{3}, \text{if } \mathrm{U_H} > \mathrm{U_M}, \mathrm{U_H} > \mathrm{U_L} \\ \frac{1-\mathrm{U_L}}{3}, \text{if } \mathrm{U_L} > \mathrm{U_M}, \mathrm{U_L} > \mathrm{U_H} \\ 0.3 \quad \text{if } \mathrm{U_L} = \mathrm{U_M} \\ 0.5 \quad \text{if } \mathrm{U_H} = \mathrm{U_M} \end{cases} \qquad (۱۳)$$



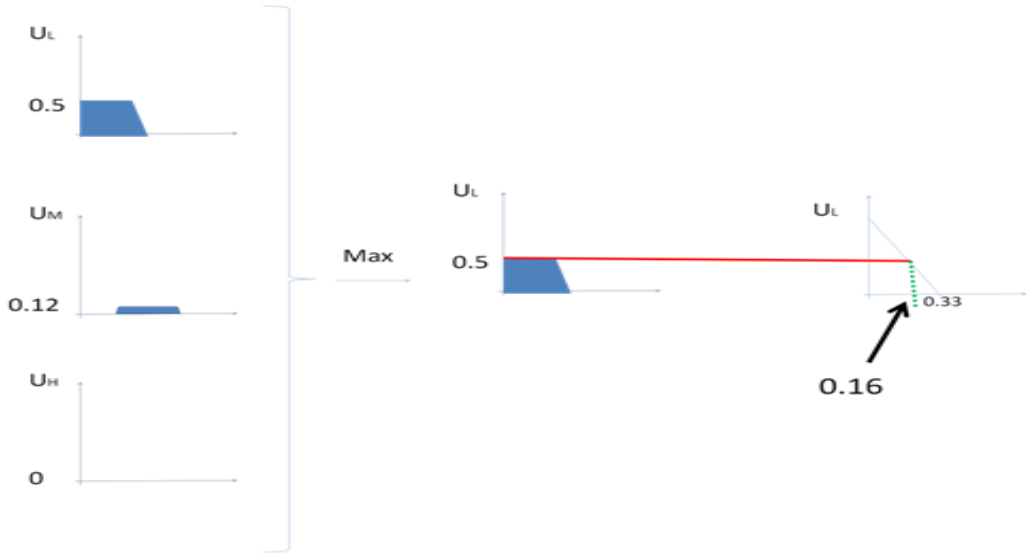
شکل ۶- روال محاسبهٔ درجهٔ عضویت به شباهت Low با ورودی‌های (۱۵٪، ۷۵٪،۲۵٪)

به عنوان مثال، فرض می‌شود میزان شباهت دو سند، بین مفاهیم اصلی مقدار ۰/۲۵، بین مفاهیم جزئی ۰/۷۵ و بین ساختار دو سند ۰/۱۵ است. برای مشخص نمودن درجهٔ عضویت شباهت بین دو سند به میزان شباهت High، میزان شباهت Medium و میزان شباهتLow، ابتدا قوانین فازی جدول۲ بررسی می‌شود. برای مشخص نمودن درجهٔ عضویت به میزان شباهتLow، قوانینی بررسی می‌شوند که میزان شباهت دو سند را Low تخمین زده‌اند. در قوانین ۱۴، ۱۵، ۱۶، ۱۷ و ۱۸ در جدول ۲، میزان شباهت Low است.

بنابراین سیستم استنتاج فازی برای تخمین درجهٔ عضویت به شباهت Low از این پنج قانون استفاده می‌نماید. شکل ۷ روال سیستم استنتاج فازی برای مشخص نمودن درجهٔ عضویت به میزان شباهت Low را نشان می‌دهد. مشابه همین روال برای مشخص نمودن درجهٔ عضویت به شباهت High و Medium نیز انجام می‌شود.

پس از مشخص نمودن درجهٔ عضویت Low، High و Medium بیش‌ترین درجه عضویت انتخاب می‌شود و عمل غیرفازی‌سازی برای تخمین میزان شباهت نهایی دو سند انجام می‌شود. شکل ۸ روال پایانی غیرفازی‌سازی را نمایش می‌دهد. همانطور که در شکل نیز مشخص است مقدار شباهت نهایی این دو سند ۱۶٪ تقریب زده شده است. ۱. خوشه‌بندی اسناد

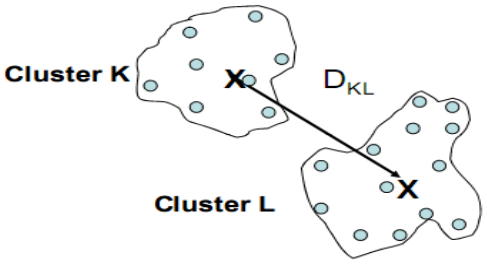
پس از محاسبهٔ ماتریس شباهت نهایی بین اسناد، با استفاده از الگوریتم خوشه‌بندی سلسله مراتبی پایین به بالا خوشه‌بندی انجام می‌شود. الگوریتم خوشه‌بندی در گام‌های زیر صورت می‌پذیرد: هر سند به عنوان یک خوشه در نظر گرفته می‌شود. ترکیب دو خوشهٔ i, j بر اساس بیش‌ترین شباهت بین این اسناد محاسبهٔ شباهت بین خوشهٔ جدید و سایر خوشه‌ها بر اساس روش centroid-linkage و ترکیب نزدیک‌ترین خوشه‌ها در ادامه، مرحلهٔ سوم تا زمانی که تنها یک خوشه باقی بماند، تکرار می‌شود.



شکل ۷- روال غیرفازی‌سازی برای ورودی‌های شکل ۷

روش centroid-linkage از فاصلهٔ بین مراکز خوشه‌ها استفاده می‌نماید. اگر i یک شی در خوشه r و j یک شی در خوشه s و n_r, n_s به ترتیب تعداد اشیاء در خوشه‌های s و r باشند، آنگاه فاصلهٔ بین دو خوشه از طریق رابطهٔ ۱۴ محاسبه می‌شود که x_{ri}، امین شی در خوشهٔ r است. شکل ۹ این اندازه‌گیری را برای دو خوشه‌ی L,K نشان می‌دهد.

$$d(r,s) = \left\| \bar{x}_r - \bar{x}_s \right\|^2 ,$$
$$\bar{x}_r = \frac{1}{n_r} \sum_{i=1}^{n_r} x_{ri} , \qquad (۱۴)$$
$$\bar{x}_s = \frac{1}{n_s} \sum_{i=1}^{n_s} x_{si}$$



شکل ۸- خوشه‌بندی سلسله مراتبی به روش Centroid

- **ارزیابی روش پیشنهادی**

الگوریتم خوشه‌بندی پیشنهادی در محیط .NET و با استفاده از C# توسعه داده شده است. همچنین از پایگاه داده رابطه‌ای MySql برای ذخیره‌سازی مفاهیم و نمونه‌ها و روابط بین آن‌ها، شمای آنتولوژیکال اسناد و ماتریس شباهت اسناد استفاده شده است. برای پیاده‌سازی سیستم استنتاج فازی در محیط .NET از چارچوب Aforge.net بهره گرفته شده است. جهت پیاده‌سازی و اجرای آنتولوژی از زبان OWL که به صورت گراف‌های جهت‌دار، روابط و منطق بین مفاهیم را تشریح می‌کند، استفاده شده است. نرم افزار به کار رفته برای پیاده‌سازی آنتولوژی، نرم افزار Protégé، محصول دانشگاه استنفورد است. این نرم‌افزار با پشتیبانی از زبان‌های RDF، OWL و قدرت تبدیل آن‌ها به یکدیگر، این امکان را فراهم می‌سازد که در صورت نیاز با استفاده از منطق توصیفی منطبق بر گزاره‌ها که OWL ارائه می‌دهد، آنتولوژی را برای استفاده در وب معنایی آماده کند.

همانطور که می‌دانیم زبانی که آنتولوژی آن را می‌شناسد OWL است چرا که مجموعه‌ای ازRDF است. در الگوریتم پیشنهادی از روش نگاشت آنتولوژی به پایگاه داده برای استخراج اطلاعات استفاده شده است. به این ترتیب که آنتولوژی از یک فایل OWL استخراج شده و

سپس در یک پایگاه داده رابطه‌ای ذخیره می‌شود. پایگاه داده رابطه‌ای اغلب به عنوان پایه‌ای برای ذخیره‌سازی آنتولوژی جهت کمک به سرعت بخشیدن به عملیاتی از قبیل جستجو و بازیابی و همچنین استفاده از مزایای سیستم مدیریت پایگاه داده رابطه‌ای مانند کنترل امنیت و یکپارچگی استفاده می‌شود. پیاده سازی طرح پیشنهاد شده به گونه‌ای صورت گرفته است که به صورت خودکار از جدول ایجاد شده از آنتولوژی، جداول کلاس‌ها و نمونه‌ها و ماتریس‌های کلاس‌ها و کلاس- نمونه ساخته می‌شود. بنابراین پیاده‌سازی انجام شده قابل تطبیق با هر آنتولوژی است. به منظور کاهش مصرف حافظه، روابط بین کلاس‌ها به صورت ماتریس کامل ذخیره نمی‌شود. تنها کلاس پدر، کلاس فرزند و تعداد روابط بین این دو کلاس ذخیره می‌شوند. برای ذخیره‌سازی ماتریس کلاس- نمونه نیز بدین صورت عمل می‌شود.

برای بررسی روش پیشنهادی، دو ارزیابی انجام شده است. در ارزیابی اول هدف خوشه‌بندی اسناد به صورت کلی و در چند دامنه است. در ارزیابی دوم خوشه‌بندی جزئی اسناد، در یک دامنه صورت می‌پذیرد. کارایی روش پیشنهادی در ارزیابی اول با چهار الگوریتم دیگر مقایسه می‌شود. الگوریتم اول، خوشه‌بندی مبتنی بر Naive Bayes است که یک خوشه‌بند کننده ساده احتمالاتی است که مبتنی بر تئوری Bayes و فرضیه‌های Naive است. در این الگوریتم هر سند با فرکانس کلمات (روش فضای بردار) نمایش داده می‌شود [۲۴]. الگوریتم دوم در مقالهٔ [۲۵] ارائه شده است. این روش از یک چارچوب مبتنی بر آنتولوژی استفاده می‌نماید. این چارچوب شامل پیش پردازش، استخراج ویژگی و کاهش ابعاد است. بخش آنتولوژی، مهم‌ترین بخش در این چارچوب است و دانش پس‌زمینهٔ مورد نظر را فراهم می‌سازد. بخش آخر بخش خوشه‌بندی نهایی اسناد است که با استفاده از آنتولوژی، مفاهیم از اسناد استخراج می‌گردند و با کمک الگوریتم خوشه‌بندی مناسب، اسناد دسته‌بندی می‌شوند.

در مقالهٔ [۲۵] برای هر سند، برداری مطابق با روش فضای برداری ساخته می‌شود. برای هر دامنه نیز برداری از مفاهیم ساخته می‌شود که به هر مفهوم با توجّه به اهمّیتش در دامنه وزنی تعلق می‌گیرد. در نهایت شباهت بین بردارهای اسناد و بردارهای ساخته شده از آنتولوژی، به صورت کسینوسی محاسبه می‌شوند. این مقاله برای ارزیابی از دویست و پنجاه سند از سه دامنهٔ مختلف پیتزا[۱۵]، نوشیدنی‌ها[۱۶] و کامپیوتر[۱۷] استفاده نموده است. صد و شش سند به دامنهٔ کامپیوتر، شصت و چهار سند به دامنهٔ پیتزا و هشتاد سند به دامنهٔ نوشیدنی متعلق است. نتایج خوشه‌بندی آن نیز با روش Bayes مقایسه شده است.

الگوریتم خوشه‌بندی سوم، روش خوشه‌بندی تکرار کننده مبتنی بر روش Bayes[۱۸] (IBC) است که از روش مزیت نسبی[۱۹]، CA برای برچسب‌گذاری اولیه اسناد استفاده می‌نماید [۲۴]. این الگوریتم با نام CA-IBC شناخته می‌شود. با توجّه به فرضیه‌های ساده روش Bayes، از این روش نمی‌توان به صورت مستقیم برای خوشه‌بندی اسناد بدون برچسب استفاده نمود. برای حل این مشکل، الگوریتم IBC مطرح می‌شود. این الگوریتم در سه گام انجام می‌پذیرد: مشخص نمودن برچسب اولیهٔ اسناد به روز رسانی برچسب همهٔ اسناد مبتنی بر روش BC پایان الگوریتم در صورت عدم تغییر برچسب اسناد در گام دوم، در غیر این صورت برگشت به گام دوم.

الگوریتم برای برچسب‌گذاری اولیهٔ اسناد از روش CA استفاده می‌نماید. روش CA فرض می‌کند هر سند یک انسان است و هر کلمه فعالیتی است که می‌تواند هر انسان انجام دهد. خوشه‌بندی اسناد روال دسته‌بندی افراد است تا سود، بیشینه گردد. الگوریتم CA ابتدا به صورت تصادفی اسناد را برچسب‌گذاری می‌نماید. سپس به خوشه‌بندی اسناد می‌پردازد. این روال تا هنگامی که برچسب اسناد تغییر کند ادامه می‌یابد. الگوریتم CA-IBC، ابتدا از طریق الگوریتم CA اسناد را برچسب‌گذاری می‌نماید و سپس

15. Pizza
16. Drink
17. Computer
18. Iterative Bayes Clustering
19. Comparative Advantage

الگوریتم IBC به خوشه‌بندی نهایی اسناد می‌پردازد. روال CA، به تعداد خوشه‌ها، اسنادی را به صورت تصادفی انتخاب می‌نماید و به عنوان بردار میانگین اولیه خوشه در نظر می‌گیرد.

الگوریتم چهارم در مقالهٔ [۲۶] ارائه شده است. در این مقاله روش خوشه‌بندی اسناد با نام CCAG[۲۰] ارائه شده است. در این روش پس از انجام پیش پردازش‌های متعارف اسناد، کلمات به مفاهیم آنتولوژی نگاشت می‌شوند. در واقع یک فضای ویژگی با ابعادی برابر با تعداد مفاهیم موجود در آنتولوژی در نظر گرفته می‌شود و هر سند در این فضا نمایش داده می‌شود.الگوریتم خوشه‌بندی ارائه شده بر روی مجموعهٔ اسناد [۲۵] اعمال می‌شود و نتیجهٔ آن با چهار الگوریتم فوق مقایسه می‌شود. در الگوریتم پیشنهادی ابتدا آنتولوژی‌های مورد نظر به پایگاه داده نگاشت ارائه می‌شوند. شکل‌های ۱۰و ۱۱ آنتولوژی‌های مورد استفاده را نشان می‌دهد.

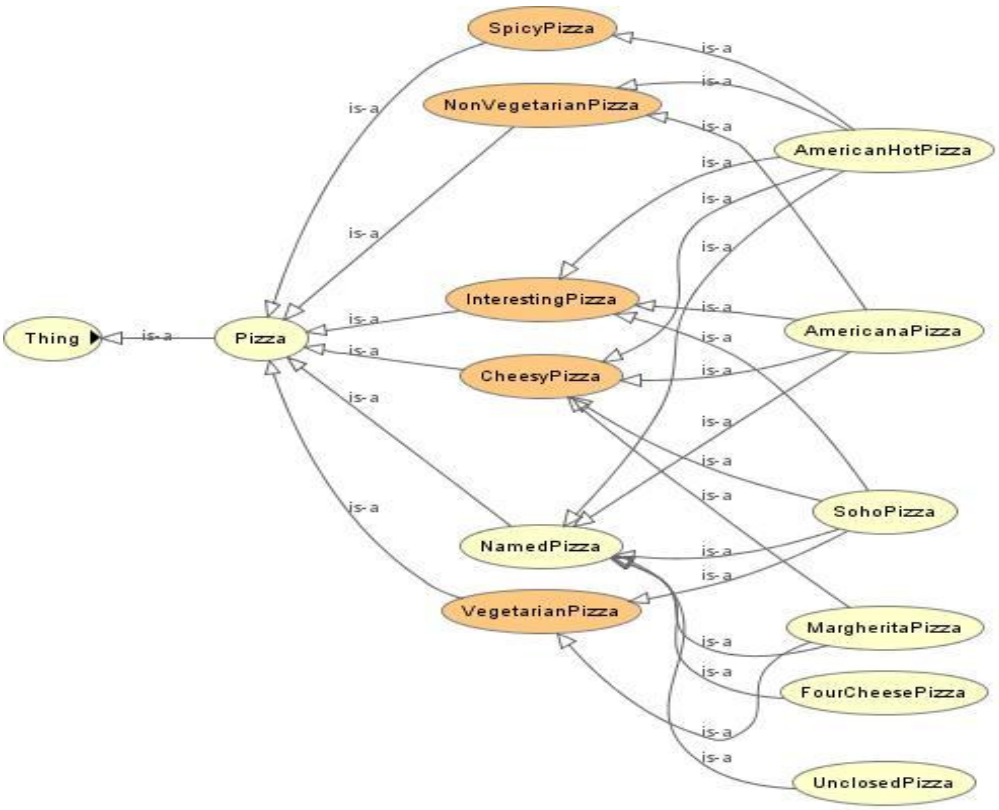
جداول و ماتریس نمونه‌ها، کلاس‌ها و روابط بین آن‌ها به صورت خودکار استخراج می‌شوند. برای پیش پردازش اسناد از حذف کلمات نویزی، نشانه‌گذاری و ریشه‌یابی استفاده شده است. نشانه‌گذار تمامی علائم نقطه‌گذاری را حذف می‌نماید و نشانه‌های صحیح را برمی‌گرداند. سپس نوبت به مرحلهٔ حذف کلمات نویزی می‌رسد. دویست و دوازه کلمه، به عنوان واژه‌های نویزی زبان انگلیسی در نظر گرفته شده‌اند و در صورت مشاهده در سند حذف می‌شوند. برای تمامی نشانه‌های باقیمانده ریشه‌یابی صورت می‌گیرد. الگوریتم مورد استفاده Porter Stemmer است. این الگوریتم برای تمامی کلمات، ریشه را به درستی تشخیص نمی‌دهد. بنابراین هر کلمه پس از ریشه‌یابی با لغات موجود در دیکشنری کاملی از زبان انگلیسی مقایسه می‌شود و به شبیه‌ترین و صحیح‌ترین کلمه نگاشت می‌شود.

پس از مراحل فوق شمای آنتولوژیکال اسناد استخراج می‌شود. برای هر مفهوم مستقیم در سند، چهار سطح مفهوم غیرمستقیم-۱، شامل فرزندان و پدران، نیز استخراج می‌شود. برای استخراج مفاهیم غیرمستقیم-۲ از دیکشنری مفهومی استفاده می‌شود. این دیکشنری از دیکشنری معکوس Onelook[۲۱] ساخته شده است. دیکشنری معکوس برای هر مفهوم مرتبط‌ترین کلمات را به ترتیب اهمّیت برمی‌گرداند. برای هر مفهوم صد کلمه بدین صورت انتخاب می‌شود. همچنین برای هر مفهوم صفحات اینترنتی مرتبط جستجو شده و کلمات متناسب با مفهوم انتخاب می‌شوند. سرانجام برای هر مفهوم کلمات یافت شده مورد بازبینی نهایی قرار می‌گیرند و مناسب‌ترین کلمات در دیکشنری مفهومی باقی می‌ماند. همانگونه که قبلاً توضیح داده شد، برای هر سند به صورت نسبی مفاهیم کلی، جزئی و یال‌های اصلی استخراج می‌شوند. برای هر یک از این اجزا سه ماتریس شباهت تولید می‌شود. برای تخمین شباهت پایانی از سیستم استنتاج فازی استفاده می‌شود. خروجی سیستم استنتاج فازی، ماتریس شباهت نهایی است که ورودی الگوریتم خوشه‌بندی سلسله مراتبی پایین به بالا است.

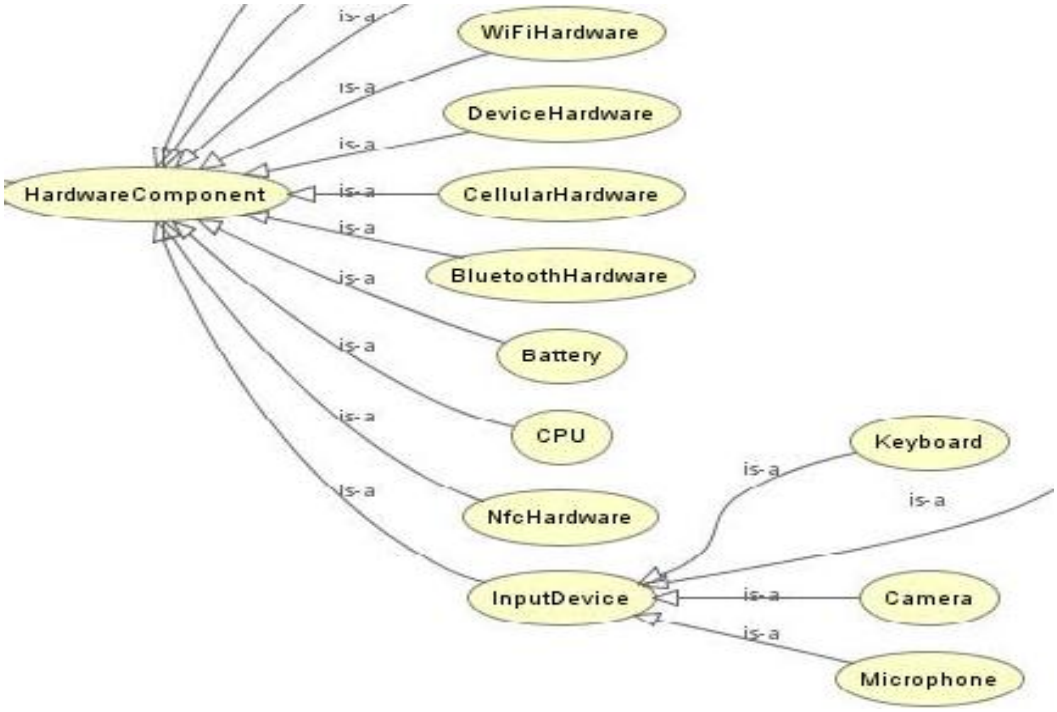
مقایسه نتایج الگوریتم پیشنهادی و دو الگوریتم دیگر از طریق شش معیار ارزیابی صورت می‌پذیرد. از معیارهای ارزیابی precision (رابطهٔ۱۵)، recall (رابطهٔ۱۶)، F-measure (رابطهٔ۱۷) و Accuracy (RI) (رابطهٔ۱۸) و همچنین از دو معیار دیگر با نام‌های Error و FP (False Positive rate) استفاده شده است که به ترتیب در روابط ۱۹ و ۲۰ بیان شده‌اند. مقدار tp، تعداد جفت متونی است که در خوشه و دستهٔ یکسانی ظاهر شده‌اند. مقاد fp, fn, tn به ترتیب false positive، false negative و true negative هستند.آزمایش‌های انجام شده برای ارزیابی الگوریتم پیشنهادی بر روی یک سیستم با ویندوز ویستا، با پردازندهٔ دو هسته‌ای ۲.۵۳ گیگا هرتز اینتل همراه با ۴ گیگا بایت حافظه، برای اسناد انگلیسی انجام گرفته است. مقدار آستانه برای رفع ابهام از یک مفهوم، به دست آوردن نمره‌ای بیش‌تر از ۸۰ است. برای هر مفهوم مستقیم تعداد چهار سطح از فرزندان و چهار سطح از پدران اضافه می‌گردد.همچنین در مرحلهٔ رفع ابهام در صورتی از مفاهیم پاراگراف‌های قبل و بعد استفاده می‌شود که نسبت مفاهیم رفع ابهام شده پاراگراف فعلی به کل مفاهیم مبهم

20 .Concept Choice And Grand Total

21 .http://www.onelook.com/reverse-dictionary.shtml

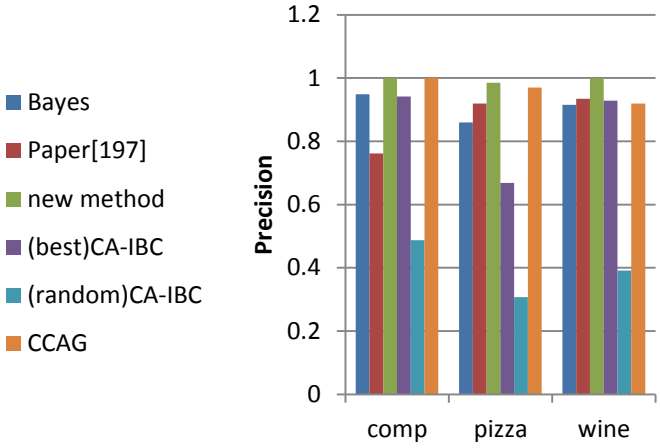


شکل ۹- بخشی از آنتولوژی پیتزا



شکل ۱۰- بخشی از آنتولوژی کامپیوتر

پاراگراف از ۱۵٪ کم‌تر باشد. برای خوشه‌بندی تمامی اسناد از ۱ تا ۲۵۰ شماره‌گذاری می‌شوند و در یک مسیر مشخص قرار می‌گیرند. همانگونه که اشاره شد از پنج الگوریتم برای خوشه‌بندی دویست پنجاه سند در سه دامنهٔ پیتزا، کامپیوتر و نوشیدنی‌ها استفاده شده است. برای خوشه‌بندی اسناد با روش CA-IBC، الگوریتم صد مرتبه با حالات برچسب‌گذاری اولیهٔ مختلف اجرا می‌شود و نتیجهٔ نهایی به صورت میانگین این نتایج اعلام می‌شود. برای این الگوریتم دو حالت در نظر گرفته شده است: ۱- روش برچسب‌گذاری اولیه کاملاً تصادفی انجام شود (random). ۲- فرض می‌شود که در صد مرتبه اجرا، اسناد مناسب به عنوان مرکز اولیهٔ خوشه در نظر گرفته شوند (best). شکل‌های ۱۲، ۱۳، ۱۴، ۱۵، ۱۶و۱۷ نتایج ارزیابی پنج الگوریتم را نشان می‌دهند. شکل ۲۰ به صورت میانگین شش معیار فوق را در پنج الگوریتم بررسی می‌کند.



شکل ۱۱- مقایسهٔ پنج روش با معیار Precision

$$p = \frac{tp}{tp + fp} \qquad (۱۵)$$

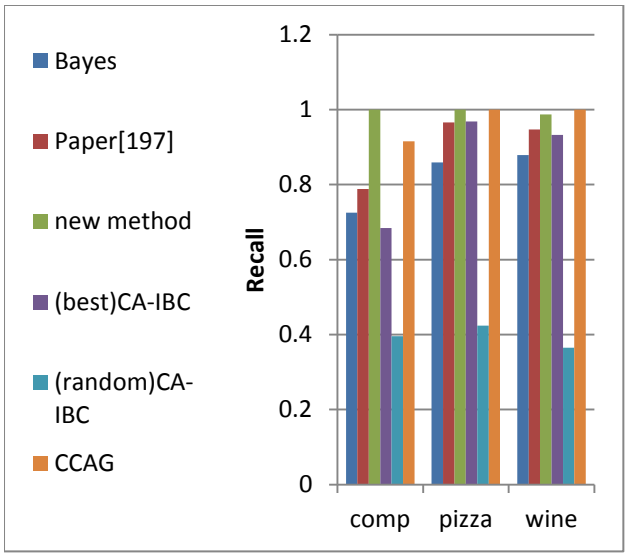
$$r = \frac{tp}{tp + fn} \qquad (۱۶)$$

$$F_\beta = \frac{(\beta^2 + 1)pr}{\beta^2 p + r} \qquad (۱۷)$$

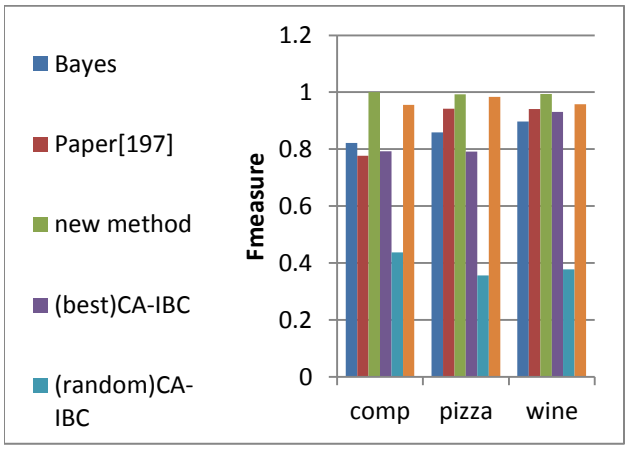
$$RI = \frac{tp + tn}{tp + fn + fp + tn} \qquad (۱۸)$$

$$Error = \frac{fp + fn}{fp + fn + tn + tp} \qquad (۱۹)$$

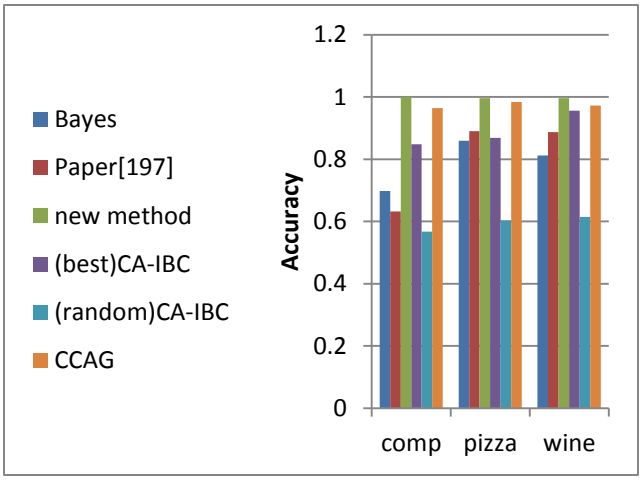
$$FP = \frac{fp}{fp + fn} \qquad (۲۰)$$



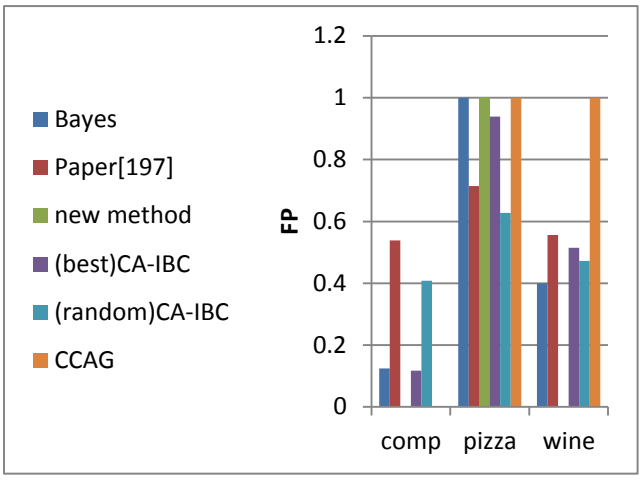
شکل ۱۲-مقایسهٔ پنج روش با معیار **Recall**



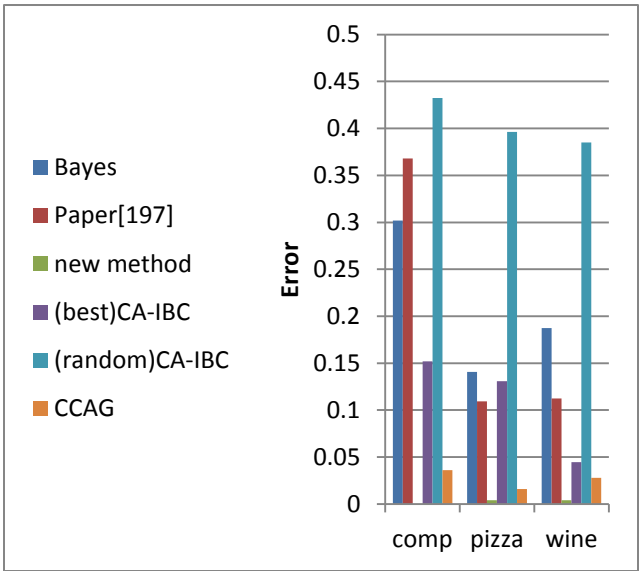
شکل ۱۳-مقایسهٔ پنج روش با معیار **Fmeasure**



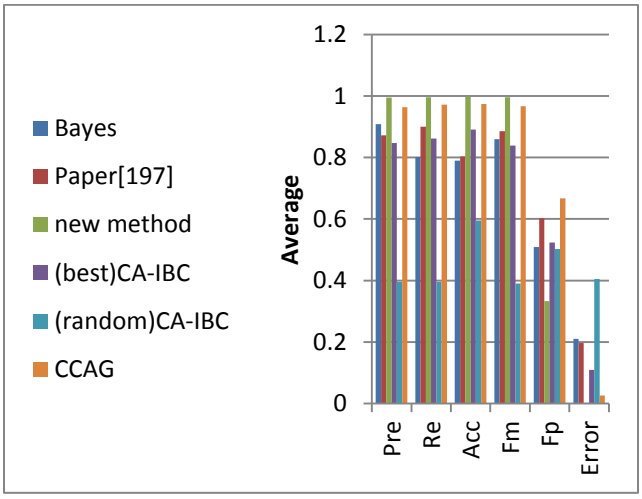
شکل ۱۴- مقایسهٔ پنج روش با معیار **Accuracy**



شکل ۱۵- مقایسهٔ پنج روش با معیار **FP**



کل ۱۶- مقایسهٔ پنج روش با معیار **Error**



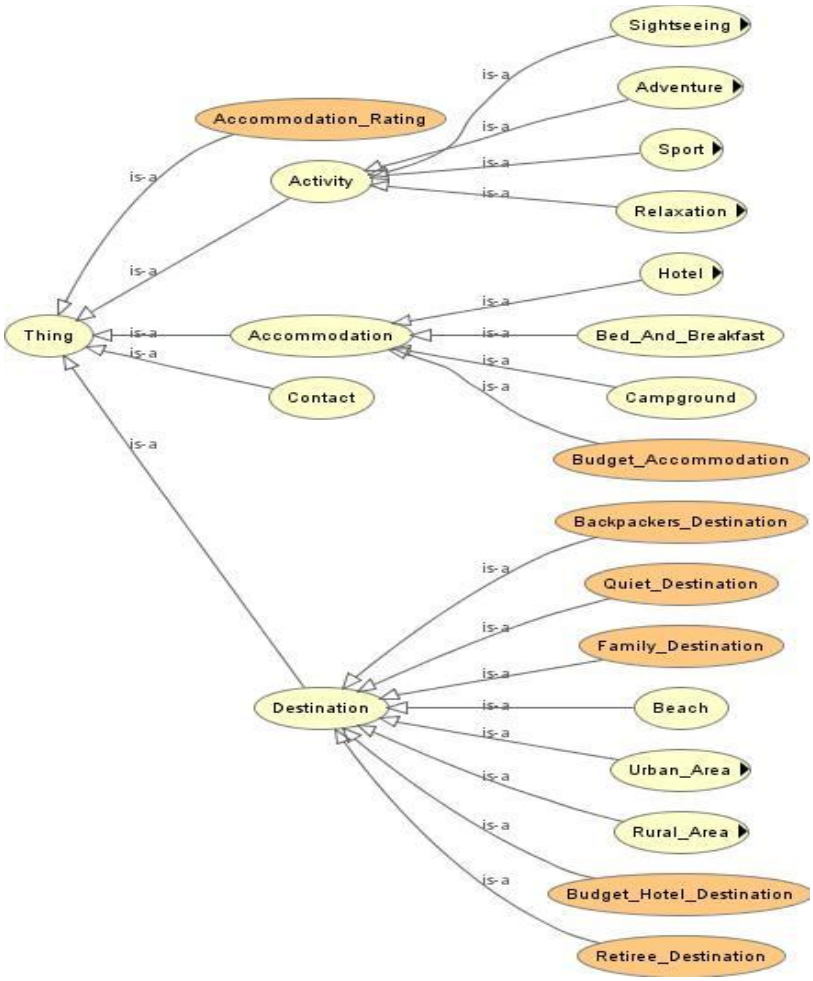
شکل ۱۷- مقایسهٔ پنج روش به صورت میانگین

همانگونه که از نتایج برمی‌آید خوشه‌بندی CA-IBC به دلیل برچسب‌گذاری تصادفی اولیه در مقایسه با روش‌های مفهومی نتایج بسیار ضعیفی تولید می‌نماید. حتی با فرض اینکه در تمامی صد بار اجرا، اسناد اولیه از خوشه‌های مناسب انتخاب شوند، نتایج تولید شده در مقایسه با روش‌های مفهومی چندان مناسب نیستند. در واقع این الگوریتم یک الگوریتم غیر قطعی است و طبق مقالهٔ [۲۴] این روش در مقایسه با سایر روش‌های آماری، با ورودی‌های یکسان، بهتر عمل می‌نماید. الگوریتم CCAG در مقایسه با روش IC-IBC بهتر عمل نموده است. روش پیشنهادی به دلیل رفع ابهام از مفاهیم و پردازش دقیق‌تر مفاهیم استخراج شده از سند، نمایش جامع‌تر و دقیق‌تر از اسناد ایجاد می‌نماید. استفاده از سیستم استنتاج فازی و ساختار مفهومی دقیق‌تر باعث می‌شود خوشه‌بندی صحیح‌تری نیز تولید شود. الگوریتم پیشنهادی تنها یک سند را به صورت نادرست خوشه‌بندی کرده است: سندی در دستهٔ نوشیدنی‌ها، وارد خوشهٔ پیتزا شده است. با بررسی مفهومی این سند مشخص شد که مضمون این سند راجع به بیماری‌های حاصل از نوشیدنی‌ها است و به مفاهیم مربوط به دامنهٔ نوشیدنی‌ها به صورت جزئی و غیرمستقیم اشاره شده است. خوشه‌بندی ایجاد شده با روش Bayes یک روش احتمالاتی است و بر کلمات تکیه دارد. با توجّه به نمایش ضعیف اسناد، نتایج خوشه‌بندی این روش نسبت به دو روش مفهومی چندان قابل توجّه نیست. روش مقالهٔ [۲۵] نیز یک روش مبتنی بر آنتولوژی است. این روش از روش فضای برداری و کاهش ابعاد برای نمایش اسناد استفاده می‌نماید. همانگونه که اشاره شد کاهش ابعاد لزوماً روابط مفهومی اسناد را استنتاج نمی‌نماید. این مقاله تنها از مفاهیم آنتولوژی برای نمایش دانش پس زمینهٔ هر خوشه استفاده نموده است و ویژگی‌ها و روابط بین مفاهیم آنتولوژی را نیز نادیده گرفته است. با فرض داشتن یک آنتولوژی جامع و کامل و همچنین یک لغت نامهٔ مفهومی مناسب می‌توان با استفاده از الگوریتم پیشنهادی، روال‌های کاوش اسناد که مبتنی بر نمایش اسناد هستند را بهبود داد. در ارزیابی دوم هدف انجام خوشه‌بندی جزئی در یک دامنهٔ خاص است. روش پیشنهادی به علّت دقت ساختار مفهومی اسناد، می‌تواند در یک دامنهٔ خاص نیز برای خوشه‌بندی استفاده شود. برای خوشه‌بندی جزئی دامنه travel در نظر گرفته شده است. بیست و پنج سند به صورت کاملاً تصادفی از این دامنه انتخاب شده است. یک سند به زیر دامنهٔ Adventure، بیست سند به زیر دامنهٔ Hotel و چهار سند به زیر دامنهٔ Sport مرتبط هستند. شکل ۱۹ بخشی از آنتولوژی travel را نمایش می‌دهد. خوشه‌بندی این مجموعه اسناد با دو روش برتر ارزیابی قبل، الگوریتم پیشنهادی و روش CCAG، انجام شده است. نتایج خوشه‌بندی با این دو روش در جدول‌های ۲ و ۳ مشاهده می‌شوند. در شکل ۲۰ نتایج حاصل از دو روش به صورت میانگین با یکدیگر مقایسه شده‌اند.

همان‌طور که از نتایج مشخص است روش پیشنهادی در مقایسه با روش CCAG، در خوشه‌بندی جزئی با موفقیت بیش‌تری عمل کرده است. نمایش دقیق مفهومی اسناد، معیار شباهت متناسب با این نمایش و سیستم استنتاج فازی باعث برتری روش پیشنهادی شده است. البته به علّت تشابه بیش‌تر ساختار مفهومی اسناد، این خوشه‌بندی کمی مشکل‌تر از خوشه‌بندی کلی است.

نتیجه‌گیری

در این مقاله یک الگوریتم خوشه‌بندی اسناد مبتنی بر نمایش آنتولوژیکال و سیستم استنتاج فازی ارائه شده است. در ابتدا مفاهیم پایه و الگوریتم‌های خوشه‌بندی اسناد، مدل‌های خوشه‌بندی، کاربردهای نمایش اسناد و سیستم‌های خوشه‌بندی اسناد مطرح شد. در ادامه الگوریتم خوشه‌بندی اسناد پیشنهادی ارائه شد. این الگوریتم در پنج مرحله صورت می‌پذیرد. در مرحلهٔ اول پیش پردازش‌های لازم بر مجموعهٔ اسناد انجام می‌شود. نشانه‌گذاری و ریشه‌یابی از مهم‌ترین پیش پردازش‌های انجام شده هستند. پس از انجام پیش‌پردازش برای هر سند مجموعه‌ای از نشانه‌های اصلی و ریشه‌هایشان نگهداری می‌شود. مرحلهٔ بعد تولید شمای آنتولوژیکال اسناد است. نمایش آنتولوژیکال پیشنهادی یک گراف وزن‌دار و جهت‌دار است که ساختار مفهومی اسناد را مشخص می‌نماید.



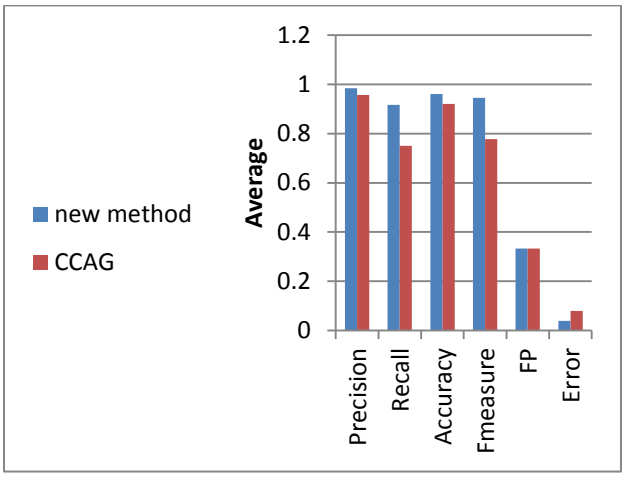
شکل ۱۸- بخشی از آنتولوژی Travel

جدول ۲- نتایج خوشه‌بندی جزئی با روش پیشنهادی

	Adventure	Sport	Hotel	Average
Precision	100%	100%	0.9523	0.9841
Recall	100%	0.75	100%	0.9166
Accuracy	100%	0.9600	0.9200	0.9600
FP	0	0	1	0.3333
Error	0	0.0400	0.0800	0.04
F1-measure	100%	0.8571	0.9755	0.9442

جدول ۳- نتایج خوشه‌بندی جزئی با روش CCAG

	Adventure	Sport	Hotel	Average
Precision	100%	100%	0.8695	0.9565
Recall	100%	0.25	100%	0.75
Accuracy	100%	0.8800	0.8800	0.9200
FP	0	0	1	0.3333
Error	0	0.1200	0.1200	0.08
F1-measure	100%	0.4000	0.9301	0.7767



شکل ۱۹- مقایسهٔ دو روش با تمامی معیارها به صورت میانگین

در واقع نمایش پیشنهادی برای هر سند زیر گرافی از آنتولوژی دامنه تولید می‌کند. در این نمایش مفهومی روال رفع ابهامی برای حذف مفاهیم مبهم در نظر گرفته شده است. این روال باعث نمایش دقیق‌تر اسناد می‌شود. گام سوم محاسبهٔ شباهت بین هر دو سند است. بر اساس نمایش آنتولوژیکال، معیار شباهت متناسبی نیز ارائه شده است. این معیار شباهت مفاهیم مشترک هر دو سند را بررسی می‌نماید و بر اساس وزن و اولویت این مفاهیم، شباهت بین دو سند را تخمین می‌زند. به منظور تخمین دقیق‌تر شباهت بین اسناد، در سه جزء بامعنی (مفاهیم اصلی، مفاهیم جزئی و یال‌های اصلی) شباهت دو سند تخمین زده می‌شود. در گام چهارم سیستم استنتاج فازی طراحی شده است که دارای سه ورودی و یک خروجی است. ورودی‌های این سیستم شباهت محاسبه شده در سه جزء و خروجی مقدار شباهت نهایی بین دو سند است. سیستم استنتاج فازی تولید شده ماتریس شباهت نهایی بین اسناد را ایجاد می‌نماید. مرحلهٔ آخر خوشه‌بندی سلسله مراتبی اسناد است که از مدل پایین به بالا با روش centroid استفاده شده است. در نهایت به ارزیابی روش پیشنهادی پرداخته شده است. نتایج حاصل از این خوشه‌بندی در مقایسه با روش‌های دیگر نشان می‌دهد که الگوریتم پیشنهادی نتایج بهتری تولید می‌نماید. موفقیت این الگوریتم نسبت به سایر الگوریتم‌ها، به دلیل استفاده از نمایش آنتولوژیکال است. این نمایش نسبت سایر روش‌های مرسوم نمایش اسناد، مضمون هر سند را به درستی درک و استخراج می‌نماید. همچنین معیار شباهت متناسب با این نوع نمایش و سیستم استنتاج فازی پیشنهادی باعث ایجاد خوشه‌بندی دقیق‌تری می‌شوند.

از مزیت‌های روش پیشنهادی، استخراج نمایش مفهومی دقیق اسناد است. این نوع نمایش همانند روش‌های قبل از یک فضای ویژگی برای همهٔ اسناد استفاده نمی‌نماید که این امر سبب کاهش حافظه مصرفی و پرهیز از مشکلات ابعاد بالا می‌شود. همچنین در نظر گرفتن اولویت مفاهیم در محاسبهٔ شباهت اسناد، نسبت به سایر معیارهای شباهت، سبب می‌شود که شباهت دو سند با دقت بیش‌تری تخمین زده شود. نمایش مفهومی و معیار شباهت ارائه شده سبب می‌شوند این الگوریتم در خوشه‌بندی جزئی نیز نتایج قابل قبولی تولید نماید. از محدّودیت‌های این روش این است که الگوریتم پیشنهادی به آنتولوژی وابسته است. برای انجام یک خوشه‌بندی دقیق باید آنتولوژی دامنه مورد نظر موجود باشد. هرچه لغت‌نامه مفهومی در آنتولوژی دقیق‌تر و جامع‌تر باشد، نتایج خوشه‌بندی با کیفیت بیش‌تری همراه خواهد بود. ناقص بودن لغت‌نامه و آنتولوژی دامنه سبب کاهش کیفیت نتایج خوشه‌بندی می‌شود. در فعالیت‌های آتی سعی در ایجاد خوشه‌بندی فازی مبتنی بر نمایش آنتولوژیکال پیشنهادی است. همچنین تعیین خودکار تعداد سطوحی که باید در آنتولوژی برای هر مفهوم بررسی شوند، به‌روزرسانی لغت‌نامه‌های مفهومی بر اساس خوشه‌های ایجاد شده در هر دامنه و انجام آزمایش‌های بیش‌تر در زمینهٔ خوشه‌بندی‌های یک دامنهٔ خاص می‌تواند مورد مطالعه و بررسی قرار گیرد.

منابع

1.G. Salton, M. J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, 1984.

2.W. Frawley, G. Piatetsky-Shapiro and C. Matheus, Knowledge Discovery in Databases: An Overview, AI Magazine, Fall 1992, pp. 213-228.

3.L.Yanjun, HIGH PERFORMANCE TEXT DOCUMENT CLUSTERING, M.S., Wright State University, 2003.

4.D.Q. Zhang and S.C. Chen, Clustering incomplete data using kernel-based fuzzy c-means algorithm, Neural Processing Letters, 18(3):155–162, 2003.

5.L. Jing, Survey of text clustering, 2006. On website: http://www.alphaminer.org/document/downloads/TextMining/survey%20of%20text%20clustering.pdf

6.M. Porter, An Algorithm for Suffix Stripping, Program, 14(3), pp. 130-137, 1980.

7.G. Salton, E.A. Fox, H. Wu, Extended Boolean information retrieval, Communications of the ACM, 26(11), pp. 1022-1036, 1983.

8.W.R. Hersch, D.L. Elliot, D.H. Hickam, S.L. Wolf, A. Molnar, C. Lechtenstien, Towards new measures of information retrieval evaluation, Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 164-170, 1995.

9.Miller, Wordnet: A lexical database for english, CACM, vol. 38, no. 11, pp. 39-41, 1995.

10.L. Khan and D. McLeod, Audio structuring and personalized retrieval using ontology, Proc. of IEEE Advances in Digital Libraries, 2000.

11.T. Gruber, A translation approach to portable ontology specifications, Knowledge Acquisition, vol. 5, no. 2, pp. 199-220, 1993.

12.Muresan, Learning to Map Text to Graph-based Meaning Representations via Grammar Induction, 3 Proceedings of the 3rd Textgraphs Workshop on Graph-Based Algorithms for Natural Language Processing, 2008.

13.P. Ljungstrand, H. Johansson, Intranet indexing using semantic document clustering. Master Thesis, Department of Informatics, Gteborg University, 1997.

14.B. Solheim,K. Vågsnes, Ontological Representation of Texts, and its Applications in Text Analysis, Agder University College,2003.

15.R.R. Korfhage, Information Storage and Retrieval, John Wiley and Sons, New York, 1997.

16.C.J. Van Rijsbergen, Information Retrieval (2nd ed.), Butterworths, London, 1979.

17.M. Rosell, M. Hassel, and V. Kann, Global evaluation of random indexing through Swedish word clustering compared to the people's dictionary of synonyms, Submitted, 2009.

18.D. Zeimpekis and E. Gallopoulos, PDDP(l): towards a flexible principal direction divisive partitioning clustering algorithm, In D. Boley, I. Dhillon, J. Ghosh, and J. Kogan, editors, Proc. Workshop on Clustering Large Data Sets (held in conjunction with the Third IEEE Int'l. Conf. Data Min.), pages 26–35, Melbourne, FL, November 2003.

19.H. Schütze and C. Silverstein, Projections for efficient document clustering, In Proc. 20th annual int. ACM SIGIR conf. on Research and development in information retrieval, pages 74–81, New York, NY, USA. ACMPress. ISBN 0-89791-836-3, 1997.

20.K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, When is nearest neighbors meaningful?, Proc. of 7th International Conference on Database Theory (ICDT'99), pp. 217- 235, 1999.

21.L. Yu and H. Liu, Feature selection for high-dimensional data: a fast correlation-based filter solution, Proc. of the 20th International Conference on Machine Learning, pp. 856-863, 2003.
22.F. Lamberti, A. Sanna, C. Demartini, A Relation-Based Page Rank Algorithm for Semantic Web Search Engines, IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 1, 2009.
23.http://aimm02.cse.ttu.edu.tw/class/92-2/ANNs/04-Apr-19-2.ppt
24.J. Ji and Q. Zhao, Applying Naive Bayes Classifier to Document Clustering, Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.14 No.6, 2010.
25.X. Yang, N. Sun, Y. Zhang, D. Kong, General Framework for Text Classification based on Domain Ontology, In SAMP 08: Proceedings of the 2008 Third International Workshop on Semantic Media Adaptation and Personalization. IEEE Computer Society. Washington DC, USA. pp. 147-152, 2008.
26.Ding. Y., Fu. X., A Text Document Clustering Method Based on Ontology, ISNN'11 Proceedings of the 8th international conference on Advances in neural networks, Volume Part II, Springer-Verlag Berlin, Heidelberg, 2011