مجله پژوهشي دانشگاه اصفهان (علوم پايه)
جلد بيست و يكم- شماره ۱- سال ۱۳۸٤
صص ۱۵٤ - ۱۳۹

موضوع بندی متن‌های متراکم

ناصر قاسم آقایی* و قاسم صرافان**

*** دانشکده فنی و مهندسی دانشگاه اصفهان**
**** موسسه آموزش عالی شهید اشرفی اصفهانی**

چکیده

در اين مقاله براي اولين‌بار پرونده‌هاي متني از لحاظ تراكم موضوعي، بررسي گرديده و به دو دسته: متن‌هاي متراكم[1] و نامتراكم تفكيک شده‌اند. متن متراكم متني است كه شامل دامنه وسيعي از موضوعات باشد، يا به عبارت ديگر ميزان چگالي موضوعي (تعداد موضوعات تقسيم بر طول پرونده) آنها بالا باشد (مانند متون مذهبي، ديوانهاي شعر، آرشيو مجلات و دايره المعارف‌ها). آنچه در اين نوشتار مورد اثبات قرار خواهد گرفت: اول عدم كارايي روشهاي موجود موضوع‌بندي، براي متن‌هاي متراكم. دوم كارايي سيستم پيشنهادي جديد، براي متن‌هاي متراكم خواهد بود.

---

۱. Dense Text

در اين تحقيق از روابـط وابـستگي، مـسيرها و پايگـاه داده‌اي از سـه‌تـايي‌هـا[1] و همچنـين روش‌هاي آماري پردازش متن، در جهت استخراج كلمات و تركيبات مهم از متن‌هاي متـراكم و قرار دادن آنها در يک سلسله مراتب خوشه‌اي، استفاده مي‌شود. همچنين در اين راستا فرآيندي بـراي يافتن مرجع ضميرها -كه خاص متن‌هاي متراكم است- تشريح خواهـد شـد. نتـايج بدسـت آمـده از نمونه پياده‌سازي شده كه بر روي متن‌هاي متراكم زبان فارسي آزمـايش شـده اسـت، نـشانگر بهبـود كيفيت انديس‌گذاري و جـستجو اسـت. همچنـين يـک سيـستم نمونـه بـه نـام نـصير[2]، بـر اسـاس روش‌هاي پيشنهادي، پياده‌سازي گرديد كه از آن براي ارزيابي نتايج بهره جسته‌ايم.

واژه‌های كليدي: پردازش متن، متن‌متراكم، موضوع بندي، روابط وابستگي، مرجع‌يابي ضمير.

Topic Specific of the Dense Documents

N. Ghasem Aghaee and G. Sarrafan

Department of Computer Engineering, University of Isfahan

Abstract

This paper investigates text documents regarding their topic density. It has divided them into two groups: dense and sparse documents. Dense documents are texts with a wide domain of topics. They have a high topic density (for example religious books, encyclopedia, magazine archives, etc). We have shown that a) traditional methods can not be used for topic

---

۱. Triples

۱- به ياد خواجه نصير الدين طوسي، حكيم، دانشمند، منجم و رياضيدان بزرگ ايراني (قرن ۷).

specific of dense texts, and b) we can benefit from employing the efficiency of the proposed method (Nasir) for dense texts.

In this research, we have used dependency relations, paths, triple databases and statistical text processing methods to extract important words and to insert them into a clustering index. Also a method was described to find the reference of pronouns in dense texts.

In addition, based on the suggested methods, a prototype system called Nasir was implemented. The result of the implementation on Persian dense texts shows that the quality of indexing and searching improved significantly.

Keywords: Text processing, Dense text, Topic specific, Dependency relations, Pronoun referencing

مقدمه

موضوع مورد تحقیق در این نوشتار، یک زمینه کاملا جدید در پردازش متن است. ما در این تحقیق براي نخستین‌بار مساله «تراکم موضوعي متن» را مطرح کرده و تأثیر آن را بر نتایج عملیات جستجو و موضوع‌بندي بررسي خواهیم کرد.

آنچه در این تحقیق محور بررسي قرار مي‌گیرد، مفهوم «تراکم موضوعي متن» و تأثیر آن بر فرایند جستجو و موضوع‌بندي متن‌ها است. در ابتدا پنج تعریف را به عنوان مفاهیم پایه این مقاله ارایه مي‌دهیم:

تعریف ۱: متن متراکم به یک پرونده متني اطلاق مي‌شود که شامل دامنه وسیعي از موضوعات باشد، یا به عبارت دیگر میزان چگالي موضوعي (تعداد موضوعات تقسیم بر طول پرونده) آنها بالا باشد (مانند متون مذهبي، دیوان‌هاي شعر، آرشیو مجلات و دایره‌المعارف‌ها).

تعریف ۲: متن نامتراکم به یک پرونده متني که داراي چند موضوع محدود هستند که با تعداد کمي کلمه مهم در متن مشخص مي‌شوند (مانند مقالات علمي، اخبار و صفحات وب).

تعريف ۳: کلمات مهم به اسامي خاص موجود در يك پرونده متني گفته مي‌شود که منعکس‌کننده موضوع يا موضوعات اصلي آن متن باشند.

تعريف ٤: کلمات عمومي به کلماتي مثل حروف ربط، اضافه، افعال معين و همچنين کلمات خاصي که به موضوع متن ارتباط ندارد اطلاق مي‌شود.

از آنجا که هدف اصلي اين مقاله بررسي روشهاي تحليل و جستجوي متن متراکم بوده است، از اينرو مشاهده الگوي پراکندگي يک نمونه که داراي حجم قابل توجهي نيز مي‌باشد براي نيل به اين هدف کافي بنظر مي‌رسد. اصولا وجود متن‌هايي با تراکم بالا و چگونگي پراکندگي کلمات مهم در آنها يک مفهوم ذهني است که بدون اتکا به مشاهدات تجربي تنها با استنتاج منطقي نيز قابل اثبات است.

در متن‌هاي عادي (نامتراکم) تعداد کلماتي که از ديدگاه موضوع‌بندي مورد توجه قرار مي‌گيرد، نسبتاً کم هستند و اين کلمات بارها در متن تکرار مي‌شوند. از سوي ديگر کلمات عمومي نيز بارها در متن تکرار شده‌اند با اين تفاوت که کلمات عمومي در متن‌هاي ديگر نيز به همين ميزان تکرار مي‌شوند. بنابراين در روش مرسوم موضوع‌بندي و انديس گذاري يعني TFIDF[۱] براي يافتن کلمات مهم يك متن از اين اصل استفاده مي‌شود و موضوع يك متن را وابسته به کلماتي مي‌دانند که در متن جاري پرشمار و در ساير متن‌ها کم شمار باشد (٤). اما آيا اين اصل براي متن‌هاي متراکم نيز صادق است؟

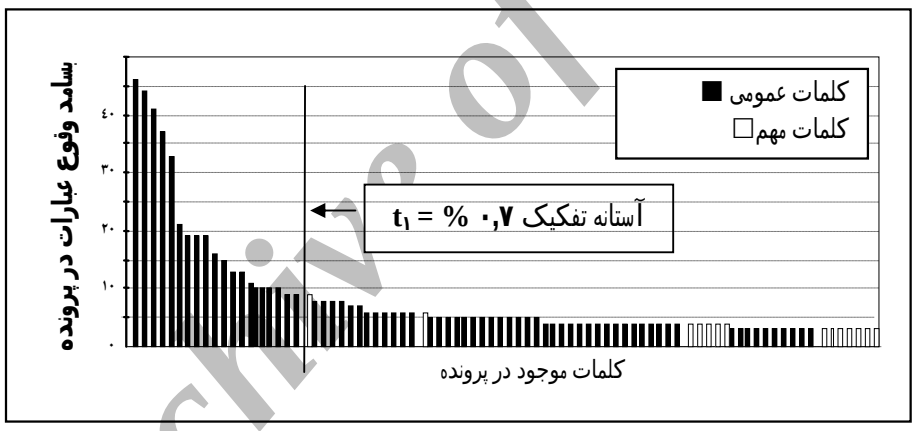
در متن‌هاي متراکم تعداد کلمات مجزايي که موضوع متن را منعکس مي‌کنند، زياد است (برخلاف متن‌هاي نامتراکم). بنابراين هر کلمه مهم در اين گونه متن‌ها با فرکانس کم مشاهده مي‌شود، به عنوان مثال اگر در دو متن، يکي متراکم و ديگري نامتراکم که هر دو داراي ۱۰۰۰ کلمه هستند، ۸۰۰ کلمه عمومي (به همراه تکرارهايشان) وجود داشته باشد، تعداد کلمات مهم هر متن و تکرارهايشان ۲۰۰ کلمه مي‌شود. اکنون اگر متن اول متراکم باشد يعني حاوي ۵۰ عنوان کلمه مهم مجزا باشد، از اينجا ميتوان نتيجه گرفت که هر کلمه مهم بطور متوسط ٤ = ۵۰ / ۲۰۰ مرتبه در متن تکرار شده است. اکنون اگر در متن نامتراکم تنها ۱۰ عنوان موضوع متفاوت وجود داشته باشد، فرکانس متوسط هر کلمه در متن ۲۰ = ۱۰ / ۲۰۰ خواهد بود. حال با نگاهي دوباره

---

۱. Term Frequency Inverse Document Frequency

به روش TFIDF که مي‌گويد: موضوعات اصلي يك متن با كلماتي مشخص مـي‌شـود كـه اول در متن جاري پرشمار باشند و دوم در ساير متن‌هاي متفاوت كم شمار باشند (٤)، نتيجـه مـي‌گيـريم كـه فرض اول در روش TFIDF با حالت متن‌هاي متراكم كه در آنها كلمات اصـلي در مـتن كـم‌شـمار مي‌باشند، منطبق نيست. بنابراين اين روش قادر به تشخيص كلمات مهم يك متن متراكم نمي‌باشد.

ما در ادامه براي درک بهتر اين واقعيت منطقي از يک نمونه آماري عيني نيز بهره برده‌ايم كه تنها جنبه تفهيمي دارد و نه اثباتي. شكلهاي يك و دو به ترتيب نـشان‌دهنـده نـوع پراكنـدگي كلمات مهم در دو نمونه از متن‌هاي متراكم و نامتراكم اسـت.(جهـت مـشاهده جـداول پايـه كـه نمودارهاي شكل يک و دو براسـاس آنهـا اسـتخراج شـده بـه (١) مراجعـه شـود.) دسـتاورد ايـن بررسي، مشخص كردن تفاوت‌هاي اساسي در نوع پراكندگي آماري، براي كلمات مهـم موجـود در اين دو نوع متن است.



شكل١: نمودار بسامد كلمات عمومي و مهم در يک متن متراكم [1]

۱- ترجمه سوره ياسين، استاد فولادوند.

شكل٢: نمودار بسامد كلمات عمومي و مهم در متن نامتراكم [1]

تعريف 0: آستانه تفكيك مرزي است كه كلمات با بسامد بيشتر از آن حتما از كلمات عمومي بشمار مي‌رود. مقدار آستانه تفكيك برابر با بيشترين درصد بسامد از ميان كلمات مهم متن است.

مقدار آستانه تفكيك (t_1) را ميتوان به عنوان فيلتر اوليه‌اي در نظر گرفت كه هر كلمه‌اي با درصد فركانس بيشتر از آن را ميتوان به عنوان يك كلمه عمومي از دامنه پردازش‌هاي بعدي حذف كرده و بار محاسباتي سيستم را به ميزان قابل توجهي كاهش داد.

در جدول يک مقاديري جهت مقايسه نشان داده شده است. مشاهدات انجام شده نشان مي‌دهد كه در اين دو نمونه سطح آستانه مناسب براي متن متراكم ٠/٧ درصد و براي متن نامتراكم ١/٢ درصد مي‌باشد. رديف‌هاي٤ و0 از جدول يك نشان مي‌دهد كه كلمات مهم در متن‌هاي متراكم (بر خلاف متن‌هاي نامتراكم) غالبا با بسامدهاي پايين مشاهده مي‌شوند.

جدول ١: مقايسه آماري ميان متن متراكم و نا متراكم

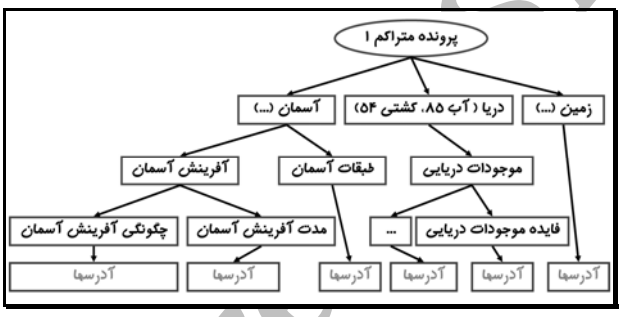
	متن متراكم	متن نا متراكم
١. تعداد كل كلمات	١٥٤٢	٢٦٧٢
٢. ميزان تراكم (تعداد موضوع در طول پرونده)	٣١/٢ درصد	٤/٣ درصد
٣. آستانه تفكيک	٧/. درصد	١/٢ درصد
٤. درصد كلمات مهم در بسامدهاي بالا	٤ درصد	١٥/٨ درصد
0. درصد كلمات مهم در بسامدهاي پايين	٥٨ درصد	٢ درصد
٦. تعداد كلمات عمومي حذف شده (بالاي آستانه تفكيک)	٤٦٦ (٣٠درصد)	١٧٠(٦/٣درصد)

١- فصل دوم از (١).

شناسايي كلمات مهم (موضوعات)

مهم‌ترين فرآيند در كل سيستم، يافتن لغات مهم يا همان موضوعات اصلي متن مـي‌باشـد. اين كلمات به تنهايي در سطح اول خوشه‌بندي قرار خواهند گرفت و با گزينش هـر يـک از آنهـا بـه عنوان موضوع مورد نظر كاربر (پژوهشگر)، مشتقات و تركيبات آن كه در سطوح بعدي قرار دارنـد، فهرست مي‌شوند.

شكل سه نمونه‌اي از يک سلسله مراتب خوشه‌بندي را نشان مي‌دهد كه در سـطح اول آن موضوعات و در سطوح بعدي زير موضـوعات قـرار مـي‌گيرنـد. اعـداد داخـل پرانتـز در قـسمت موضوعات، ميزان ارتباط معنايي (مترادف بودن) موضوع را با ساير موضوعات مشخص مي‌كند.



شكل ۳ : خوشه بندي موضوعات و زير موضوعات در يک متن متراكم

تعيين كلمات مهم حتما بايد پس از فرآيند مرجع‌يابي ضماير انجام پذيرد. هر چند پس از انجام اين فرآيند باز هم ممكن است ضمايري بدون مرجع بمانند و الگوريتم، قادر به يافتن مرجع آنها نشود. در اين صورت مانند كلمات عمومي، از آنها چشم‌پوشي مي‌شود.

براي تعيين كلمات مهم علاوه بر شناسايي آنها از طريق يک فرهنگ لغت سـاختاريافته[۱]، ميتـوان از مشاهده ميزان فراواني آنها در كل متن متراكم نيز به وجود آنهـا پـي بـرد. عموميـت (عـدم اهميـت) عبارتي مثل t با $generality(t)$ نشان داده مي‌شود. كلمـاتي عمـومي بـه شـمار مـي رونـد كـه در همـه پرونده‌ها پرشمار باشند. رابطه ۱ متوسط تعداد وقوع عبارت را در همه پرونده ها انـدازه‌گيـري مـي‌كنـد.

۱. Thesaurus

هنگامي نتيجه رابطه عدد بزرگي مي‌شود كه يک عبارت به طور متوسط در همه پرونده‌ها پرشمار ظاهر شود. براي هر عبارت t ميتوان نوشت(٤):

$$\text{generality(t)} = \frac{\sum_{d}^{N} \frac{f_d(t)}{L_d}}{N} \qquad ١$$

در اين رابطه N تعداد پرونده‌هاي محيط، L_d طول پروندۀ d (تعداد كل كلمات متن) و $f_d(t)$ ميزان بسامد عبارت t در پرونده d مي‌باشد. اين رابطه را ميتوان با در نظر گرفتن ضريب ميزان تراكم هر پرونده به شكل زير بهبود بخشيد و نتيجه‌ی دقيقتري حاصل كرد. چون در متن‌هاي نامتراكم كلمات مهم نيز با بسامد بالا وجود دارند، ضريب تراكم پايين آنها باعث كاهش تاثير اين خطا و جلوگيري از حذف بي‌مورد كلمات مهم به جاي كلمات عمومي مي‌شود (١):

$$\text{generality(t)} = \frac{\sum_{d}^{N} \frac{f_d(t).density(d)}{L_d}}{N} \qquad ٢$$

در اينجا $density(d)$ كه ميزان تراكم پرونده d را مشخص مي‌كند، بدين شكل محاسبه مي‌شود (١):

$$density(d) = \frac{N_{imp\_term}(d)}{L_d} \qquad ٣$$

N_{imp_term} تعداد كلمات مهم يا به عبارتي تعداد موضوعات پرونده d را با تساحب بسامد آنها، مشخص مي‌كند (١):

$$N_{imp\_term}(d) = \sum_{i}^{L_d} f_{t_i} | w(t_i) > w\_threshold$$

$density(d)$ براي شروع، مقادير اوليه‌ی $generality$ را از رابطه‌ی ١ بدست مي‌آورد و مي‌تواند پس از آن بصورت تكرار شونده[١] با مقادير رابطه‌ی ٢ خود را تصحيح كند. اين پردازش‌ها زمانگير است و در صورت تمايل به تسريع فرآيند، ميتوان تنها به رابطه‌ی ١ بسنده كرد. مقادير $generality$ مشخص كننده وزن هر عبارت نيز هست (١):

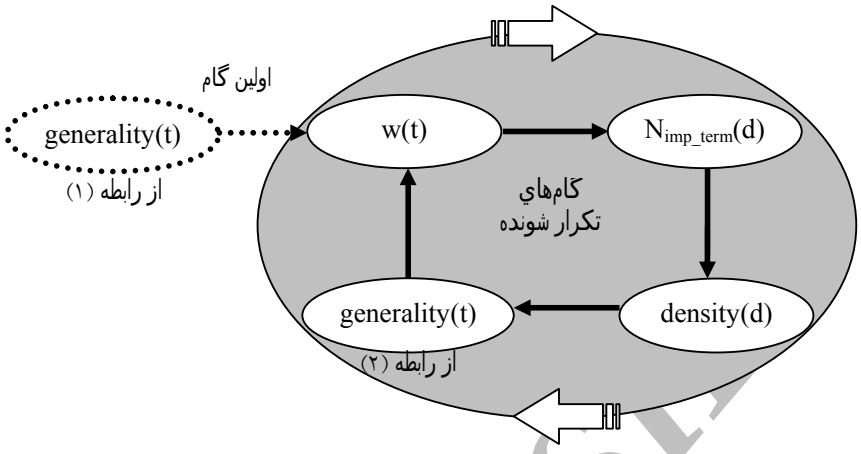
١. Iterative

$$w(t) = \frac{۱}{generality(t)} \qquad ٤$$

كلماتي كه ميزان w(t) [۱] بدست آمده‌ی آنها كمتر از سطح آستانه‌ی w_threshold باشد در خوشه‌بندي موضوعات ناديده گرفته خواهند شد. براي متن‌هاي حساس‌تر ميتوان با افزايش مقدار w_threshold با احتياط بيشتري عمل نموده و دامنه‌ی كلمات مهم را وسيع‌تر نمود. ميزان حد آستانه بصورت پويا و بر اساس بازخورد جستجوهاي قبلي و امتياز بندي نتايج توسط موتورهاي جستجو تعيين و به تدريج متعادل مي‌شود. انتخاب حد آستانه كوچكتر، تعدادكلمات مهم بيشتري را به عنوان موضوع متن انتخاب مي‌كند. برعكس حد آستانه بزرگتر مي‌تواند دقيق‌تر و باحساسيت بيشتري تعدادكمتري كلمه را بازگرداند كه اين امر در فرايند جستجو باعث افزايش درصد پاسخ‌هاي مربوط مي‌شود.

طريقه‌ی استفاده از روابط فوق بدين شكل است كه ابتدا generality تمام عبارات براي پرونده d توسط رابطه ۱ محاسبه مي‌شود. مقادير بدست آمده هر چند دقيق نبوده و بدون توجه به چگالي موضوعي متن بدست آمده است، با اين وجود به عنوان مقدار تقريبي اوليه‌اي براي محاسبه، چگالي كاربرد دارد. با استفاده از چگالي تقريبي بدست آمده اين‌بار generality جديد و دقيق‌تري با استفاده از رابطه ۲ محاسبه مي‌شود. با تكرار مراحل اين فرآيند generality ها در هر گام دقيق‌تر مي‌شوند و حلقه تا جايي ادامه مي يابد كه اختلاف generality هاي بدست آمده در دو گام آخر نزديك به صفر شود. شكل چهار ترتيب استفاده از روابط را نشان مي‌دهد:

۱- براي توضيح ساده‌ترعملكرد روابط، ابتدا generality(t) محاسبه گرديده و بر اساس آن w(t) بدست مي‌آيد.



شكل ٤ : ترتيب اجراي تكرار شونده روابط و تاثير چگالي موضوعي متن در محاسبه وزن كلمات

با انتخاب كليد واژه *t* توسط پژوهشگردر زمان جستجوي اطلاعات، بازخورد اين انتخاب براي تقويت مقدار *w(t)* به صورت خودكار، بكار مي‌رود. افزايش اين مقدار مي‌تواند بعضي عبارات را از ليست كلمات عمومي خارج كرده و در حقيقت خطاي محاسبات اوليه را جبران كند. همچنين سيستم، اين اجازه را به كاربر مي‌دهد تا مستقيما وزن عبارات را تنظيم كند.

در اينجا Δw مقداري است كه در صورت وجود عبارت *t* در مجموعه كليد واژه‌هاي تايپ شده توسط كاربر، به *w(t)* آن افزوده مي‌شود، مثلاً اگر وزن عبارت t برابر۲ باشد و كاربر از آن كلمه در جستجو استفاده كند از آن پس وزن عبارت t برابر $\Delta w + ۲$ قرار مي‌گيرد (۱):

$$\Delta w = \frac{w\_threshold}{۲ \times (\log_{۲}(N_{keyword} + ۱))} \qquad ۵$$

$N_{keyword}$ تعداد كلمات كليدي انتخاب شده در يک درخواست كاربر مي‌باشد. مقدار Δw نيز براساس بازخورد سيستم به صورت پويا و با افزايش پايه لگاريتم قابل تنظيم است.

شناسايي تركيبات مهم (زيرموضوعات)

پس از شناختن کلمات برجسته متن و قرار دادن آنها در سطح اول خوشه‌بندي (موضوعات)، نوبت به شناسايي تركيبات مهم و قرار دادن هر يک از آنها در سطوح بعدي خوشه‌بندي مي‌رسد. يک تركيب مي‌تواند داراي دو يا چند عبارت باشد و لازم نيست كه همه‌ی عبارات يک تركيب كلماتي مهم باشند و گاهي تنها وجود يک كلمه‌ی مهم كافي است تا تركيب، براي قرار گرفتن در خوشه‌بندي قابليت پيدا كند. تركيباتي مهم به شمار مي‌آيند كه حداقل حاوي يک كلمه مهم باشند. بنابراين براي بيان تعلق تركيب مهم‌اي مثل $set = [t1 , t2 , \dots , tn]$ به موضوع (يا عبارت) ti ميتوان نوشت:

$$\forall ti \in Set \mid imp\_term(ti) \longrightarrow child (set , ti) \qquad ٦$$

يعني تركيب Set (مثل «چگونگي آفرينش آسمان ») به عنوان فرزند عبارت ti در خوشه‌بندي بكار مي‌رود، اگر ti از موضوعات برجسته‌ی متن (مثل «آفرينش » و «آسمان») و عضو تركيب Set باشد.

فرآيند مرجع‌يابي ضمایر

اين فرآيند را ميتوان مخصوص متن‌هاي متراكم دانست زيرا در متن‌هاي نامتراكم و در روش‌هاي آماري معمولا ضميرها نقش مؤثري ايفا نمي‌كنند و به راحتي ميتوان از آنها به عنوان كلمات غير اطلاعاتي، چشم‌پوشي نمود، اما با ديدن مثال‌هايي از متن‌هاي متراكم به اين نتيجه خواهيم رسيد كه نمي‌توان به سادگي از ضمير موجود در يک جمله صرف‌نظر كرد.

اگر اسامي خاص براساس ترتيب وقوع در متن درون يک پشته[1] ذخيره شوند، ميتوان ضمير را به n اسم بالاي پشته (مراجع نامزد) ارجاع داد. دقيق‌ترين روش كشف مرجع براي ضميرها تطبيق n اسم آخر(مراجع نامزد) با صافي‌هاي[2] موجود در پايگاه داد سه‌تايي است. پايگاه داده‌ی سه‌تايي محل ذخيره سه‌تايي‌هايي به شكل (P, Slot, w) است كه در آن P مشخصه يك مسير استخراج شده از جملات متن و Slot مشخص‌كننده يك شكاف (مانند يك متغير در برنامه‌نويسي) در مسير P است كه مي‌تواند يك اسم خاص را در نقش‌هاي فاعل، مفعول، صفت و ... بپذيرد و w كلمه‌ايست

---

۱. Stack

۲. Filters

که در Slot قرار گرفته است. در ادامه همین بخش یـك مثـال کـه نـشان‌دهنـده روابـط سـه‌گانـه وابستگي، مسير و شكاف مي‌باشد بيان شده است. (براي توضيحات بيشتر به (۲) رجوع شود.)

مـسير اسـتخراج شـده از يـک جملـه کـه حـاوي ضـمير اسـت در پايگـاه داده جـستجو شـده و صافي‌هاي شكاف[۱] متناظر با ضمير (شكاف X يا شكاف Y) بررسـي مـي‌شـوند(۱). در صـورتي کـه يكي از كلمات نامزد جزو صافي‌ها باشد، به عنوان مرجع ضـمير قـرار مـي‌گيـرد و چنانچـه دو يـا سـه كلمه‌ی نامزد در صافي‌ها موجود باشند آنگاه براساس ميزان وابستگي هـر صـافي بـا شـكاف مـورد نظـر (اطلاعات متقابل) كه از رابطه ۷ محاسبه مي‌شود (۲)، كلمات نامزد بصورت نزولي رتبه‌بندي مي‌شوند:

$$mi(p, Slot, w) = \log\left(\frac{|\,p, Slot, w\,| \times |\,*, Slot, *\,|}{|\,p, Slot, *\,| \times |\,*, Slot, w\,|}\right) \qquad ۷$$

سپس كلمه‌اي كه داراي بيشترين ميزان وابستگي باشد، از ميان آنها برگزيده مـي‌شـود. در مثال زير، از جمله‌ی «و او شما را از خاک آفريد»، مسير p قابل استخراج است:

«و او شما را از خاک آفريد.»
مسيرP : N : فاعل : V ← از خاک آفريد → V : مفعول : N
معناي مسير: X ، Y را از خاک آفريد.

سه‌تايي زير در برگيرنده‌ی ضمير است، كه اينجا در نقش فاعل ظاهر مي‌شود:

(P , SlotX , «او»)

اگر مسيرP در پايگاه داده موجود باشد، آنگاه صافي‌هاي مربوط به شكاف X بررسي مي‌شوند و يک ليست از اسامي بازيابي مي‌شوند، كه يكي از آنها كلمـه‌ی «خـدا» اسـت. اگـر ايـن كلمـه در جملات قبلي واقع شده باشد و در نتيجه جزو ليست مراجع نامزد باشد، در ايـن صـورت بـه جـاي

۱. Slot

کلمه‌ی «او» مي‌نشيند و در خوشه‌بندي آدرس‌دهي مي‌شود. بديهي است كه درصورت عدم وجود كلمه مورد نظر، مترادف‌هاي آن نيز بررسي مي‌شوند، تا احتمال مطابقت افزايش يابد.

ارزيابي كارايي سيستم

ميزان كارآيي سيستم با استفاده از مشاهده‌ی نتايج آن و مقايسه با نتايج يک سيستم جستجوي مرسوم قابل اندازه‌گيري است. اين كارآيي از دو جهت قابل بررسي است: «دقت»[1] و «فراخواني مجدد»[2]. پارامتر دقت، مشخص‌كننده‌ی درصد پاسخ‌هاي صحيح سيستم نسبت به كل پاسخ‌هاي صحيح ممكن است و فراخواني، درصد پاسخ‌هاي مربوط را نسبت به كل پاسخ‌ها مشخص مي‌كند.

براساس نظريات و روابط بيان شده، يك نرم افزار موضوع‌بندي به نام «نصير» پياده‌سازي گرديد، (نرم‌افزار و كدهاي آن در ديسك فشرده ضميمه (۱) موجود است) كه به كمك يك فرهنگ لغت ساختاريافته كه ريشه افعال و اسم يا صفت‌بودن كلمات و نيز مترادف‌هاي هر كلمه را مشخص مي‌كند، قادر به موضوع‌بندي و جستجوي متن مي‌باشد.

در اين ارزيابي به عنوان متن نمونه ترجمه جزء ۳۰ قرآن كريم كه داراي تراكم موضوعي بالايي مي‌باشد (براساس جدول يک برابر ۳۱/۲ درصد) ، با روش‌هاي زير مورد جستجو قرار داديم. براي هر جستجو تنها از يك كليد واژه استفاده شده است و از كليد واژه‌هاي تركيبي صرف‌نظر كرده‌ايم.

۱- **جستجوي واژه‌اي محض**: در اين روش كليد واژه تنها به صورت رشته‌اي از كاراكترها در متن جستجو مي‌شود و در صورت يافتن رشته‌اي كه تمام يا بخشي از آن مطابق با كليد واژه باشد، آدرس آن را بر مي‌گرداند (مثل عملكرد Find در اكثر نرم‌افزارها) هردو پارامتر «دقت» و «فراخواني مجدد» در اين روش به دليل غير آگاهانه‌بودن آن و عدم توجه به هيچ اطلاعات راهنمايي كننده‌اي، پايين است.

---

۱. Precision

۲. Recall

۲- جستجو در درخت ترکیبات دست‌ساز: آزمون این بخش را با استفاده از نـرم‌افـزار قرآني «هدي» نسخه ۱ انجام داده‌ایم که در آن با استفاده از فیش برداري موضـوعي یـك درخت اندیس‌گذاري شده در سه سطح موضوعات اصـلي، موضـوعات فرعـي و آدرس گزینه‌ها ایجاد گردیده است. به دلیـل اینکـه ایـن انـدیس‌گـذاري توسـط کارشناسـان بـه صورت دستي انجام گرفته، همه پاسخ‌هاي بازیابي شـده مربـوط بـه موضـوع بـوده و ایـن روش از «فراخواني مجدد»۱۰۰درصد برخـوردار اسـت و درسـت بـه همـین دلیـل حجـم آدرس‌دهي‌هاي موجود نسبت به کل جوابهاي صحیح پایین است (میزان «دقت» کم است). در حالیکه با استفاده از یك الگوریتم تحلیـل مناسـب توسـط ماشـین بـسیاري از آنهـا قابـل دسترسي مي‌باشند.

۳- جستجوي کلید واژه‌اي در سیـستم «نـصیر» نـسخه ۱ : در ایـن نـرم‌افـزار بـراي جستجوي واژه‌اي تنها به یك جستجوي رشته‌اي ساده بسنده نشده‌است، بلکـه بـا افـزودن اطلاعات یك فرهنگ لغت ساختاریافته که نوع کلمات، مترادف‌ها و ریـشه افعـال را در خود نگه مي‌دارد، فرآیند جستجوي واژه‌اي با تبدیل به یك روش آگاهانه بهبود یافته است. در این نرم‌افزار با جستجوي یك کلید واژه، مترادف‌هاي آن کلمه و اشـکال مختلـف زمان‌هاي افعال نیز براي آن قابل شناسایي و بازیابي است. این امر باعث افـزایش قابـل توجـه پارامتر «دقـت» مـي‌شـود. هرچنـد «فراخـواني مجـدد» بـه دلیـل نـاقص‌بـودن فرهنـگ لغـت ساختاریافته و سادگي فرایند تحلیل در این نمونه پیاده‌سازي شده، همچنان ضعیف است.

٤- در این روش با استفاده از قواعـد تولیـد ترکیبـات موضـوعي، یـك درخـت موضـوع‌بنـدي ماشیني در دو سطح ایجاد گردید که امکان جستجوي اندیسي را براي کاربر ایجاد مي‌کند.

با بررسي مـشاهدات نتـایج تجربـي سیـستم بـراي دو نـوع جـستجوي ممکـن یعنـي روش جستجوي واژه‌اي و نیز جستجوي ترکیبات اندیس‌گذاري‌شده، نتایجي بدست آمـد کـه در جـدول دو نشان داده شده‌اند.

جدول ۲ : مقايسه ميزان كارآيي نمونه‌ی پياده‌سازي‌شده (نصير۱) با روش‌هاي موجود

	روش	دقت	فراخواني مجدد
	جستجوي واژه‌اي محض	۵۲ درصد	۳٤ درصد
	جستجو در درخت ترکیبات دست‌ساز[1]	۳۲ درصد	۱۰۰ درصد
نصـــــير (نسخه۱)	جستجوي واژه‌اي	۷۵ درصد	٤۰ درصد
	جستجو در درخت ترکیبات ماشيني	٤۵ درصد	٦٤ درصد

دقت نسبتاً بالاي جستجوي واژه‌اي در سيستم نصير بدليل استفاده از ترادف، هم‌خانوادگي و عمليات يكپارچه‌سازي لغات، حاصل شده است. اين فرآيندها همچنين بر روي ميزان فراخواني اين روش تأثيرگذار است.

اصولاً از روش جستجوي درختي بدليل دارابودن فراخواني بهتر نسبت به روش جستجوي واژه‌اي براي مواردي كه به پاسخ‌هاي كاملا مربوط نياز باشد، استفاده مي‌شود. مشكل اين روش كم بودن دقت و در نتيجه از دست دادن درصد بيشتري از اطلاعات نسبت به روش جستجوي واژه‌ايست. بنابراين اگر بتوان دقت يا به عبارتي فراواني آدرسهاي بازگشتي را تا حد روش واژه‌اي افزايش داد كارآيي اين روش بسيار زياد خواهد شد. همانطور كه اشاره شد انجام اين كار توسط نيروي انساني بسيار پرهزينه است. اما روش خودكار پيشنهادي اين قابليت را به صورت بالقوه دارد كه به اين هدف نايل گردد.

هر چند نتايج حاصل از نمونه‌ی پياده‌سازي‌شده بخصوص از نظر دقت چندان جالب توجه به نظر نمي‌رسد، اما با شمارشهاي دستي مشخص شد چنانچه مكانيزم توليد تركيبات مصدري از افعال پياده‌سازي مي‌شد ميزان دقت سيستم به حدود ۷۰ درصد مي‌رسيد. (مثل كشف تركيب «آفرينش انسان» از جمله‌ی «انسان را از گل خشكيده‌اي آفريد.»)

نتايج

استفاده از روش‌هاي تحليل زبان هرچند كه نسبت به روش‌هاي آماري پيچيده‌تر است ولي مشاهدات نشان مي‌دهد كه مي‌توانند به نتايج دقيق‌تري منجر شود(۳). استفاده از روش

۱- براساس انديس گذاري موضوعي نرم افزار فرآني «هدي۱»

انديس‌گذاري سه‌تايي، علاوه بر كاهش پيچيدگي پردازش زبان طبيعي، باعث افزايش چشمگير كارايي نسبت به روش‌هاي مرسوم كه غالبا مبتني بر نظريه كيف كلمات هستند، خواهد شد.(۱)

روش استفاده شده در سيستم نصير تركيبي از يک روش آماري و يک روش تحليلي است. روش آماري گونه‌اي از روشTFIDF است (٤) كه براي استفاده در متن‌هاي متراكم با نوع رفتار آماري كلمات مهم در آنها، سازگار شده است. از اين روش به منظور وزن‌دهي كلمات متن و تشخيص كلمات عمومي و مهم استفاده مي‌شود. در كنار اين روش از يک روش تحليلي به نام مسيرهاي وابستگي استفاده شده است. استفاده از روابط وابستگي (۲) ميان كلمات واقع در جمله از آنجا كه به كلمات به صورت مستقل نمي‌نگرد و سعي در درک مفهوم متن با استفاده از روابط بين آنها دارد، براي هدف ما كه موضوع‌بندي متن متراكم است، مناسب مي‌باشد.

براساس اين روش و با كمک روابط آماري ارايه شده، ميتوان كلمات مهمي كه منعكس‌كننده موضوعات اصلي پرونده متني هستند را براي قرار گرفتن در راس خوشه‌هاي موضوعي، شناسايي نمود. همچنين تركيبات مختلف حاصل از چند عبارت كه بايد در سطوح بعدي خوشه‌بندي قرار گيرند، از اين كلمات مهم مشتق مي‌شوند.

يافتن مرجع ضميرها در متن باعث بهبود نتيجه فرآيند مي‌شود، زيرا مرجع يک ضمير كه معمولا از اسامي خاص مي‌باشد ممكن است جزو كلمات مهم يا به عبارتي جزو موضوعات يک جمله باشد. همچنين با بكارگيري يک شبكه معنايي ميان كلمات و استخراج قوانين استنتاجي ميتوان نتايج سيستم را به نتايج انساني نزديک‌تر نمود.

لازمه‌ي پياده‌سازي مكانيزم توليد تركيبات مصدري از افعال، داشتن پايگاه اطلاعاتي مؤثر از روابط وابستگي است. توليد اين اطلاعات بايد توسط يک تجزيه‌گر[۱] زبان مانند Minipar[۲] انجام پذيرد. نقطه ضعف اساسي اين روش وابستگي دقت آن با دقت تجزيه‌گر جملات است، كه درخت وابستگي را توليد مي‌كند. نتايج نامطلوب در تجزيه جملات زبان مي‌تواند كارايي نهايي سيستم را به شكل قابل توجهي كاهش دهد.

---

۱. Parser

۲. Minipar: www.cs.ualberta.ca/~lindek/minipar.htm.

منابع

۱- ق. صرافان، "طراحي يک سيستم هوشمند موضوع بندي و تعيين ارتباطات معنايي در متن‌هاي متـراکم فارسي" ، پايان‌نامه کارشناسي ارشد نرم‌افزار، گروه کامپيوتر دانشگاه اصفهان، ۱۳۸۲.

۲. D. Lin and P. Pantel, "DIRT: Discovery of Inference Rules from Text," In Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, ۲۰۰۱.

۳.J. J. Lin, "Indexing and Retrieving Natural Language Using Ternary Expressions", Master's Thesis, Massachusetts Institute of Technology, ۲۰۰۱.

٤.Kjersti Aas and Line Eikvil, "Text Categorization: a Survey", www. citeseer.nj.nec.com, June ۱۹۹۹.