سیزدهمین کنفرانس ملی انجمن کامپیوتر ایران
جزیره کیش، خلیج فارس، ایران ۱۹ الی ۲۱ اسفند ۱۳۸۶
انجمن کامپیوتر ایران
Computer Society of Iran

دانشگاه صنعتی شریف
دانشکده مهندسی کامپیوتر

مقاله کامل

استخراج کلمات کلیدی اسناد فارسی

مسلم محمدی جنقرا
عضو هیئت علمی
دانشگاه آزاد اسلامی واحد ملکان
mo_mohammadi@comp.iust.ac.ir

مرتضی آنالویی
استادیار دانشکده مهندسی کامپیوتر
دانشگاه علم و صنعت ایران
analoui@iust.ac.ir

چکیده: این مقاله، یک روش آماری ترکیبی، برای استخراج کلمات کلیدی۱ اسناد فارسی، پیشنهاد کرده است. روش پیشنهادی مبتنی بر پیکره متنی۲ می‌باشد. ابتدا عمل ریشه‌یابی و حذف کلمات عمومی۳انجام می‌گیرد. سپس ویژگی‌های آماری برای کلمات مختلف محاسبه شده و با استفاده از فازی‌سازی و اعمال قواعد فازی، کلمات کلیدی محتمل، انتخاب می‌شوند. گام بعدی محاسبه رخداد همزمان۴ پیشین و پسین کلمات کلیدی محتمل، با کلمات تکرار شونده،۵ در جملات سند است. با اعمال یک آستانه وفقی۶ روی رخداد همزمان کلمات، کلمات کلیدی دو کلمه‌ای را مشخص می‌کنیم. بر خلاف اکثر روش‌های آماری که فقط کلمات کلیدی یک کلمه‌ای را استخراج می‌کنند، استفاده از این روش کلمات کلیدی دو کلمه‌ای نیز استخراج می‌شوند. استفاده از ترکیب روش فازی و رخداد همزمان کلمات بهبود خوبی را نشان می‌دهد و کلمات کلیدی بامعنی‌تری پیشنهاد می‌کند.

کلمات کلیدی: استخراج کلمات کلیدی، اسناد فارسی، رخداد همزمان، فازی

1- مقدمه

با گسترش روزافزون رسانه های ذخیره سازی الکترونیکی و رسانه های ارتباطی، اطلاعات زیاد و جامعی در دسترس هستند. این اطلاعات می‌توانند به صورت فایل های تصویری، فایل های صوتی، اسناد الکترونیکی یا اخبار ارسال شده از یک گروه خبری باشند. با گسترش اسناد الکترونیکی، پیدا کردن اسناد مورد نظر از بین حجم عظیمی از اطلاعات متنی به صورت دستی کاری دشوار و در عمل غیر ممکن خواهد بود. یک راه حل برای این مشکل استفاده از کلمات کلیدی می‌باشد.

هر عبارت یا کلمه‌ی مهمی که محتویات داخل سند را تشریح کند،کلمه کلیدی گفته می‌شود. به عبارت دیگر کلمات کلیدی نشانگر خلاصه‌ای کوتاه از محتویات داخل سند می‌باشند. استخراج کلمات کلیدی یک تکنیک مهم برای بازیابی اسناد، صفحات وب، دسته‌بندی اسناد، خلاصه‌سازی، استخراج متن و غیره می‌باشد. با استفاده از استخراج کلمات کلیدی مناسب می‌توانیم سند مورد نظر خود را از بین اسناد مربوطه انتخاب کنیم.

روش‌های استخراج کلمات کلیدی از جنبه های مختلف مد نظر قرار می‌گیرد. یک تقسیم بندی به صورت زیر انجام می‌گیرد:

- اگر مجموعه ای از اسناد، با کلمات کلیدی مشخص برای هر کدام وجود داشته باشد، فرایند استخراج کلمه کلیدی یک یادگیری باناظر خواهد بود. در غیر این صورت بدون ناظر خواهد بود.
- استخراج کلمه کلیدی می‌تواند بر پایه مجموعه ای از اسناد یا یک سند باشد.
- استخراج کلمه کلیدی می‌تواند با استفاده از دیکشنری باشد. این باید مورد توجه قرار گیرد که هدف استفاده از دیکشنری ریشه‌یابی و استخراج کلمات مترادف هر کلمه می‌باشد.
- استفاده از تکنیک پردازش زبان طبیعی[7] و تحلیل بخش های گفتار.

تقسیم‌بندی دیگر برای روش‌های استخراج کلمه‌های کلیدی می‌تواند به شکل زیر باشد:

- روشهای آماری: مبتنی بر تحلیل فراوانی کلمه‌ها.
- روشهای نحوی: مبتنی بر تجزیه زبانی[8] و انطباق الگو.
- روشهای ساختاری: بررسی عنوان و رئوس کلی مطالب سند.
- روشهای ادراکی: مبتنی بر استفاده از پایگاه دانش برای تفسیر معنی و مفهوم.

هدف این مقاله استخراج کلمات کلیدی مبتنی بر پیکره، برای متون فارسی با استفاده از روش آماری می‌باشد. ابتدا فراوانی کلمه در سند و فراوانی کلمه در کل اسناد محاسبه می‌شود، سپس تعداد اسناد شامل کلمه مورد نظر به دست می‌آید و با استفاده از روش فازی کلمات کلیدی محتمل استخراج می‌شوند و در مرحله بعد

رخداد همزمان کلمات پرتکرار و کلمات محتمل، محاسبه شده و کلمات کلیدی نهایی استخراج می‌شوند.

2- کارهای مرتبط

TF*IDF یکی از پرکاربرد ترین روابط در حوزه بازیابی اطلاعات متنی می‌باشد [1,2]. که از حاصلضرب فراوانی کلمه در فراوانی معکوس سند به دست می‌آید. این روش یک روش مبتنی بر چند سند می‌باشد. [1] یک روش مبتنی بر TF*IDF برای استخراج واژه‌های کلیدی اسناد و متون فارسی ارائه کرده است که در بخش نتیجه‌گیری نتایج این روش با روش پیشنهادی مقایسه خواهد شد.

در [3] یک روش مستقل از زبان برای زبان های ژاپنی و انگلیسی مطرح شده است که با استفاده از تحلیل ساختواژی[9] و استخراج عبارات اسمی و دسته‌بندی آنها، کلمات کلیدی مناسب را فقط با استفاده از یک سند استخراج می‌کند. که بخش ساختواژی وابسته به زبان هست. و هر زبان ساختواژی مخصوص به خود را دارد.

در [4] یک الگوریتم استخراج کلمات کلیدی برای زبان انگلیسی پیشنهاد شده است، که روی یک سند اعمال می شود. ابتدا کلمات تکرار شونده استخراج می شوند و سپس مجموعه‌ای از رخدادهای همزمان هر کلمه با کلمات تکرار شونده ایجاد می‌شود. توزیع رخدادهای همزمان، اهمیت کلمه در سند را نشان می دهد. اگر توزیع احتمال رخدادهای همزمان بین کلمه a و کلمات تکرار شونده به یک زیر مجموعه خاص از کلمات تکرار شونده بایاس شود کلمه کلیدی بودن a محتمل است.این الگوریتم کارآیی قابل مقایسه با TFIDF دارد.

[5] کل کلمات را به سه دسته‌ی کلمات عمومی، کلمات خاص[10] و کلمات کلیدی تقسیم می‌کند و مرز این کلمات را به صورت فازی تعیین می‌کند. این روش نیز برای زبان انگلیسی ارائه شده است که مبتنی بر چند کلاس و چند سند می‌باشد. ایده‌ی [5] این است که کلماتی که در کلیه اسناد کلیه کلاس‌ها پخش شده باشند، دارای ارزش و مفهوم کمتری هستند و به عنوان کلمات عمومی شناخته می‌شوند. کلماتی که فقط در کلیه اسناد یک کلاس پخش شده باشند و فراوانی زیادی داشته باشند و در سایر کلاس‌ها فراوانی کمی داشته باشند به عنوان کلمات خاص یا اصطلاحات فنی[11] آن کلاس در نظر گرفته می‌شوند. و کلماتی را که در یک سند خاص دارای فراوانی زیاد بوده و در بقیه دارای فراوانی کمی باشند، به عنوان کلمات کلیدی آن سند در نظرگرفته می‌شوند.

3- استخراج کلمات کلیدی

استخراج کلمات کلیدی یکی از مباحثی هست که در آن عدم قطعیت وجود دارد یعنی تمام کلمات سند می‌توانند به عنوان کلمه‌ی کلیدی کاندید شوند. تعیین مرز بین کلمات کلیدی و غیر کلیدی کاملاً واضح نیست. بنابراین می‌توانیم از یک روش فازی برای تعیین مرز استفاده کنیم. در ادامه، مراحل استخراج کلمات کلیدی به ترتیب تشریح می‌شود.

3-1- کلمات عمومی فارسی

بعضی از کلمات در همه‌ی متون با فراوانی زیاد وجود دارند که ارزش محتوایی ندارند، مثل ضمایر، قیود، حروف اضافه و ربط و بعضی از افعال پرتکرار. به این کلمات، کلمات عمومی گفته می‌شود. با حذف کلمات عمومی در متن‌کاوی[12] آماری میزان محاسبات کم شده و کارآیی روش‌ها نیز بیشتر می‌شود. ما از یک لیست کلمات عمومی مطابق جدول 1 و 2 پیشنهاد شده در [6] استفاده کرده‌ایم.

3-2- ریشه‌یابی

یکی از مهمترین کارها در استخراج کلمات کلیدی از متون فارسی، ریشه یابی کلمات می‌باشد. هدف از ریشه‌یابی حذف اضافات از کلمه و رسیدن به ریشه‌ی اصلی کلمه هست. روش‌های مختلفی برای ریشه یابی کلمات فارسی پیشنهاد شده است[7,8] . روشی که در این مقاله استفاده شده است یک روش مبتنی بر حذف پسوندها و پیشوندها می‌باشد. این روش خیلی شبیه به روش porter [9] در زبان انگلیسی می‌باشد. با این تفاوت که روش porter فقط حذف پسوندها را انجام می‌دهد، و ما پیشوندها را نیز حذف می‌کنیم.

3-3- استخراج ویژگی‌ها و تعیین کلمات کلیدی

ابتدا مطابق جدول 1و2 کلمات عمومی حذف می‌شوند، سپس کلمات، ریشه‌یابی شده و دوباره کلمات عمومی تولید شده، حذف می‌شوند. مرحله بعدی ایجاد فرهنگ لغات (T) می‌باشد. فرهنگ لغات مجموعه‌ای از کلمات هست که تمامی کلمات موجود در اسناد را پوشش می‌دهد. به عبارت دیگر هر کلمه‌ای که حداقل یک بار در

جدول (1)- لیست کلمات عمومی (حروف پرتکرار)

در	نیز	برای	یا	را
به	تا	ها	دو	های
از	ما	آن	آنها	و
که	باید	وی	اما	نمی
این	اند	یک	دیگر	هر
با	هم	خود	اگر	ای
می	همچنین	بر		

. **جدول(2) - لیست کلمات عمومی (افعال پرتکرار**

گویم	داشتیم	توانستند	بود	باید	است
گویند	شد	توانستیم	بودم	بتوان	آمد
گویی	شده	توانم	بودن	بتواند	آمدم
گویید	شود	توانند	بودن	بتوانم	آمدن
گوییم	کرد	توانی	بوده	بتوانی	آمدند
گیرد	کردم	توانید	بودی	بتوانیم	آمده
گیرم	کردن	توانیم	بودید	بتوانید	آمدی
گیرند	کردند	خواست	بودیم	بتوانند	آمدید
گیری	کرده	خواستم	بیا	بخواه	آمدیم
گیرید	کردی	خواستن	بیاب	بخواهم	آورد
گیریم	کردید	خواستند	بیابد	بخواهد	آوردم
می‌شود	کردیم	خواسته	بیابم	بخواهند	آوردن
هست	کن	خواستی	بیابند	بخواهی	آوردند
هستم	کند	خواستید	بیابی	بخواهید	آورده
هستند	کنم	خواستیم	بیابید	بخواهیم	آوردی
هستی	کنند	خواهد	بیابیم	بکن	آوردید
هستید	کنی	خواهم	بیاور	بکند	آوردیم
هستیم	کنید	خواهند	بیاورد	بکنم	آورم
یابد	کنیم	خواهی	بیاورم	بکنند	آورند
یابم	گرفت	خواهید	بیاورند	بکنی	آوری
یابند	گرفتم	خواهیم	بیاوری	بکنید	آورید
یابی	گرفتن	داد	بیاورید	بکنیم	آوریم
یابید	گرفتند	دار	بیاوریم	بگو	آید
یابیم	گرفته	دارد	بیاید	بگوید	آیم
یافت	گرفتی	دارم	بیایم	بگویم	آیند
یافتم	گرفتید	دارند	بیایند	بگویند	آیی
یافتن	گرفتیم	داری	بیایی	بگویی	آیید
یلفتند	گفت	دارید	بیایید	بگویید	آییم
یافته	گفتم	داریم	بیاییم	بگوییم	باش
یافتی	گفتن	داشت	تواند	بگیر	باشد
یافتید	گفتند	داشتم	توانست	بگیرد	باشند
یافتیم	گفته	داشتن	توانستم	بگیرم	باشی
	گفتی	داشتند	توانستن	بگیرند	باشم
	گفتید	داشته	توانستند	بگیری	باشید
	گفتیم	داشتی	توانسته	بگیرید	باشیم
	گوید	داشتید	توانستی	بگیریم	

مجموعه اسناد ظاهر شده باشد، در فرهنگ لغات قرار می‌گیرد . t_k نشانگر کلمه kام در فرهنگ لغات می‌باشد.

پس از انجام پیش پردازش روی اسناد، ماتریس فراوانی کلمات در اسناد TF مطابق رابطه‌ی (1) ایجاد می‌گردد.

$$TF(d_i,t_k)=\frac{freq(d_i,t_k)}{N(d_i)} \quad \begin{matrix} k=1,2,...,|T| \\ i=1,2,...,N \end{matrix} \qquad (1)$$

$Freq(d_i,t_k)$: فراوانی کلمه‌ی kام در سند iام.

$N(d_i)$: تعداد کل کلمات موجود در سند iام.

$|T|$: تعداد کلمات موجود در فرهنگ لغات.

N: تعداد کل اسناد موجود.

ماتریس فراوانی کلمات در اسناد، یک ماتریس N* |T| خواهد بود. بعد از ایجاد ماتریس TF، بردار فراوانی جمعی کلمات TTF با استفاده از رابطه‌ی (2) ایجاد می‌شود.

$$TTF(t_k)=freq(t_k) \qquad (2)$$

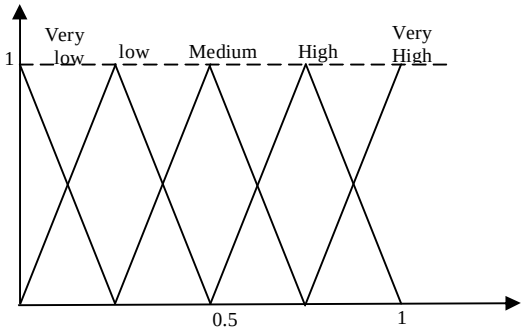
$freq(t_k)$: فراوانی کلمه kام در کل اسناد.

برای مشخص کردن موجودیت کلمات در اسناد مختلف از بردار فراوانی اسناد (DF استفاده می‌کنیم که طبق رابطه‌ی (3) تعریف می‌شود.DF در بازه‌ی [1/N,1] خواهد بود.

$$DF(t_k)=\frac{N(t_k)}{N} \qquad (3)$$

$N(t_k)$: تعداد اسناد شامل کلمه‌ی kام.

چون مقادیر ماتریس TF و بردار TTF در یک بازه‌ی نرمال نیستند، به نحوی نرمال سازی می‌شوند که در بازه [0,1] قرار گیرند. برای انجام این کار از تقسیم مقادیر بر بیشترین مقدار موجود در ماتریس و بردار، استفاده کرده‌ایم.



شکل(1) : تابع تعیین درجه عضویت

بعد از محاسبه و نرمال‌سازی، مقادیر ماتریس TF و بردارهای DF و TTF، با استفاده از یک فازی‌ساز با پنج عبارت زبانی مطابق شکل 1 فازی‌سازی می‌شوند. ماتریس و بردارهای فازی‌سازی شده را با $\overline{TF}$ و $\overline{DF}$ و $\overline{TTF}$ نشان می‌دهیم. هر متغیر زبانی شامل یک عبارت زبانی و یک درجه عضویت به صورت روابط (4) و (5) می‌باشد.

$$L_{\overline{TF}(d_i,t_k)} \in \{very\ \ low, low, medium, high, very\ \ high\} \quad (4)$$

$$m_{\overline{TF}(d_i,t_k)} \in [0,1] \quad (5)$$

مرحله بعدی اعمال قواعد فازی برای استخراج کلمات کلیدی محتمل هست. در این مقاله از 16 قاعده‌ی فازی استفاده شده است که بعضی از قوانین در جدول 3 آورده شده است. در این مرحله 10 کلمه که دارای درجه عضویت بیشتری باشند به عنوان کلمات کلیدی محتمل انتخاب می‌شوند. این کلمات را با K_p نشان می‌دهیم. تا اینجا یک لیستی از کلمات کلیدی را در اختیار داریم ولی تمام این کلمات کلیدی، یک کلمه‌ای هستند. با استخراج کلمات پرتکرار(F) در هر سند، که برای هر سند 10 کلمه پرتکرار را انتخاب می‌کنیم، و ترکیب با کلمات کلیدی محتمل(K_p) یک لیست به دست می‌آوریم که در مراحل بعدی برای استخراج کلمات کلیدی دو کلمه‌ای استفاده می‌شود. این لیست را با FK_p نشان می‌دهیم.

از دیگر مراحل این مقاله تشخیص و جداسازی جمله‌ها می‌باشد. هدف از جداسازی جمله‌ها، محاسبه رخداد همزمان لیست F و لیست FK_p هست. در این مقاله برای تشخیص جمله‌ها، از سه علامت انتهای جمله یعنی "؟"، "!" و "." استفاده شده است. چون بعد از حذف کلمات پرتکرار، احتمال وجود جمله‌های خالی یا با تعداد کلمات کم وجود دارد از یک آستانه برای طول جمله استفاده شده است. که در اینجا این آستانه 3 در نظر گرفته می‌شود.

کلمات کلیدی در اکثر مقالات و اسناد به صورت یک کلمه‌ای، دو کلمه‌ای و یا بیشتر مشخص می‌شوند. مشکلی که در بعضی از روش‌های رایج [5] استخراج کلمات کلیدی وجود دارد، استخراج کلمات کلیدی یک کلمه‌ای هست، و کلمات کلیدی دو کلمه‌ای یا بیشتر را استخراج نمی‌کنند. ما در این مقاله با استفاده از محاسبه رخداد همزمان پسین و پیشین F و FK_p، در جملات مختلف سند، بر اساس روابط 6 و 7 ، کلمات کلیدی دو کلمه‌ای را نیز استخراج می‌کنیم.

$$pre\_occ(t_j, t_f) = \sum_{i=1}^{|S|} \sum_{z=1}^{|S_i|} occ(S_i(z-1), t_j) \quad (6)$$

$$post\_occ(t_j, t_f) = \sum_{i=1}^{|S|} \sum_{z=1}^{|S_i|} occ(S_i(z+1), t_j) \quad (7)$$

$$occ(S_i(z-1), t_j) = \begin{cases} 1 & if\ (S_i(z) == t_f \ \ \& \\ & S_i(z-1) == t_j) \\ 0 & o.w \end{cases} \quad (8)$$

t_j : کلمه‌ی jام از لیست FK_p

t_f : کلمه‌ی fام از لیست کلمات پرتکرار(F)

$S_i(z)$: واژه zام در جمله‌ی iام از سند.

$|S|$: تعداد جملات هر سند.

بعد از محاسبه رخداد همزمان پسین و پیشین، با اعمال یک آستانه، که در این مقاله از یک آستانه‌گیر وفقی استفاده شده است، کلمات کلیدی، دو کلمه‌ای را استخراج می‌کنیم.

جدول3- قواعد فازی اعمال شده

Fuzzy variable / Rule number	$L_{\overline{TF}(d_i,t_k)}$	$L_{\overline{DF}(t_k)}$	$L_{\overline{TTF}(t_k)}$	$m(keyword(d_i,t_k))$
1	VH	M	M	$\min(m_{\overline{TF}(d_i,t_k)}, m_{\overline{DF}(t_k)}, m_{\overline{TTF}(t_k)})$
2	VH	L	M	$\min(m_{\overline{TF}(d_i,t_k)}, m_{\overline{DF}(t_k)}, m_{\overline{TTF}(t_k)})$
3	VH	VL	M	$\min(m_{\overline{TF}(d_i,t_k)}, m_{\overline{DF}(t_k)}, m_{\overline{TTF}(t_k)})$
4	H	VL	L	$\min(m_{\overline{TF}(d_i,t_k)}, m_{\overline{DF}(t_k)}, m_{\overline{TTF}(t_k)})$
5	M	VL	VL	$\min(m_{\overline{TF}(d_i,t_k)}, m_{\overline{DF}(t_k)}, m_{\overline{TTF}(t_k)})$
6	L	VL	VL	$\min(m_{\overline{TF}(d_i,t_k)}, m_{\overline{DF}(t_k)}, m_{\overline{TTF}(t_k)})$

$$fk(d_i,k)=\begin{cases} t_j+t_f & pre\_occ(t_j,t_f)>th_i \ \ \& \\ & pre\_occ(t_j,t_f)>post\_occ(t_j,t_f) \\ t_f+t_j & post\_occ(t_j,t_f)>th_i \ \ \& \\ & post\_occ(t_j,t_f)>pre\_occ(t_j,t_f) \end{cases} \quad (9)$$

$$th=\frac{N(d_i)}{300} \quad (10)$$

عدد 300 در مخرج کسر، با آزمایش و خطا بدست آمده است. اگر در اسناد کلمات کلیدی دو کلمه‌ای وجود نداشته باشد، یا کم باشد از کلمات کلیدی یک کلمه‌ای پیشنهاد شده توسط روش فازی نیز استفاده می‌کنیم.

4- نتایج تجربی

مجموعه‌ای که برای ارزیابی روش پیشنهاد شده در این مقاله استفاده شده است، بخشی از مجموعه متن های « محک»[10] می‌باشد. این مجموعه که از خبرگزاری ها جمع‌آوری شده است شامل اخبار و مقالاتی در اندازه‌ی نیم صفحه تا چندین صفحه می‌باشد. « محک» شامل 3007 سند، 216 پرس و جو در مورد آنها و لیست اسناد مرتبط با این پرس و جوها می‌باشد. برای ارزیابی روش پیشنهادی از معیارهای بازخوانی[13] و دقت[14] [11,12] استفاده شده است.

$$recall=\frac{\#\{relevant \cap retrieved\}}{\#\ relevant} \quad (10)$$

$$precision=\frac{\#\{relevant \cap retrieved\}}{\#\ retrieved} \quad (11)$$

با انتخاب اسناد و استخراج کلمات کلیدی، پرس‌وجوهای ذکر شده را روی آنها اعمال کرده و معیارهای بازخوانی و دقت را محاسبه کردیم. نتیجه کار و مقایسه آن با نتایج نمایه‌ساز سینا در جدول 4 آورده شده است.

جدول 4- نتایج ارزیابی

	بازخوانی	دقت در روش پیشنهادی	دقت در نمایه ساز سینا
	1	0.54	0.475
	0.9	0.57	0.532
	0.8	0.63	0.547
	0.7	0.72	0.625
	0.6	0.76	0.701
	0.5	0.85	0.758
	0.4	0.91	0.835
میانگین	---	0.71	0.6685

همانطور که در جدول 4 نیز مشاهده می‌شود میانگین دقت روش پیشنهادی 0.71 می‌باشد که نسبت به دقت نمایه‌ساز سینا بهبود یافته است. از دلایل این بهبود دقت می‌توان به استفاده از روش فازی و همچنین استخراج واژه‌های کلیدی دو واژه‌ای اشاره کرد.

5- نتیجه‌گیری و کارهای آینده

ما در این مقاله از یک روش ترکیبی برای استخراج کلمات کلیدی فارسی استفاده کردیم. ابتدا با استفاده از روش فازی کلمات کلیدی محتمل را استخراج کرده و سپس با محاسبه رخداد همزمان پسین و پیشین، کلمات کلیدی دو کلمه‌ای را استخراج کردیم. از مزیتهای روش ما سرعت بالا نسبت به روش TF*IDF و استخراج کلمات کلیدی دو کلمه‌ای می‌باشد.

یکی از مشکلاتی که در حوزه بازیابی اطلاعات متنی فارسی وجود دارد، تعیین محدوده کلمات می‌باشد. کلماتی وجود دارند که از دو یا سه بخش جدا از هم تشکیل شده‌اند. مثلاً کلمه «بین المللی» در روش‌های آماری دو کلمه در نظر گرفته می‌شود. اگر ابتدا محدوده کلمات را تعیین کنیم و سپس کلمات کلیدی را استخراج کنیم نتایج بهتری حاصل می‌شود.

مراجع

[1] بشیری، حسن، کربلائی، فاطمه، موسوی، شیرین، **طراحی و ارزیابی نمایه‌ساز خودکار متون فارسی**. مجموعه مقالات یازدهمین کنفرانس بین‌المللی کامپیوتر، بهمن 1384.

[2] G. Salton, C. Buckley. "*Term weighting approaches in automatic text retrieval*". Information Processing & Management, 24(5):513–523, 1988.

[3] D. Bracewell, F. Ren and S. Kuroiwa. "*Multilingual Single Document Keyword Extraction for Information Retrieval*", Proceedings of 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE'05), pp.517-522, Wuhan, Oct. 2005.

[4] Y. Matsuo, M. Ishizuka, "*Keyword Extraction from a Single Document using Word Co_occurrence Statistical Information*," Proceedings of the Sixteenth International Florida Artificial Intelligence Research Society Conference, pp. 392-396, 2003.

[5] M. Makrehchi, M. Kamel, "*A fuzzy set approach to extracting keywords from abstracts*", North American Fuzzy Information Processing Society- NAFIPS 2003, Banf, Canada, 2004

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

[7] K. Taghva, R. Beckley and M. Sadeh. "*A Stemming Algorithmfor the Farsi Language*". In proceedings of International Conference on Information Technology:

Coding and Computing (ITXX05) - Volume I pp. 158-162.

[8] A. Mokhtaripour, S. Jahanpour. "*Introduction to a new Farsi stemmer*". CIKM 2006: 826-827.

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

[10] K. Sheykh Esmaili, H. Abolhassani, M. Neshati, E. Behrangi, A. Rostami and M. Mohammadi Nasiri "*Mahak: A Test Collection for Evaluation of Farsi Information Retrieval Systems*", To appear in Proceedings of 5th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA-07), Amman, Jordan, May 2007.

[11] G. Salton, M.J Mc Gill, "*Introduction to Modern Information Retrieval*", Mc Graw Hill, New York, 1983

[12] R. Baeza-Yates, B. Ribeiro-Neto, "*Modern Information Retrieval*", ACM Press, 1999.

زیرنویس‌ها

[1] keywords
[2] corpus
[3] stopwords
[4] Co_occurrence
[5] frequent
[6] adaptive
[7] Natural language processing
[8] linguistic
[9] morphology
[10] feature
[11] terminology
[12] Text mining
[13] recall
[14] precision