شیوه جداسازی کلیدواژه‌ها از مدارک فرامتنی [1]

بن چویی، باولین لی[2]
ترجمه فرشید دانش[3]

چکیده

این مقاله شیوه‌ای برای جداسازی کلیدواژه‌های مدارک متنی یا فرامتنی ارائه می‌نماید. این کلیدواژه‌های جدا شده، همانند کلیدواژه‌های فهرست شده در یک مقاله، محتوای مدرک را مشخص می‌کند. برای مثال، می‌توان از فرایند پیشنهادی برای نشان دادن محتوای مدارکی که به زبان اچ. تی. ام. ال[4] از موتورهای جستجو بازیابی می‌شوند، استفاده کرد.. این شیوه به کاربران اجازه می‌دهد که اطلاعات مورد نیاز خود را به سرعت پیدا کنند. این شیوه نه تنها همانند شیوه‌های دیگر، به بسامد واژه‌های موجود در مدرک توجه دارد بلکه بسامد مترادف‌های آن واژه را نیز مدنظر قرار می‌دهد. با استفاده از روش یادشده عبارات کلیدی شامل دو یا سه واژه نیز بررسی می‌شوند. دراین شیوه برای افزایش درستی بسامد شمارش واژه‌ها، الگوریتم ریشه‌یابی برای حذف پسوند واژه‌ها به‌کار برده می‌شود. آزمایش‌ها نشان می‌دهد که به‌طور متوسط ۵۶/۷ درصد از مجموع زمان محاسبه شده، در استفاده از الگوریتم ریشه‌یابی صرف می‌شود، همچنین فرایند پیشنهادی توانسته به‌طور میانگین ۵۲ درصد از کلیدواژه‌های تهیه شده توسط مؤلفان مدارک مورد آزمایش را جداسازی نماید.

کلیدواژه‌ها

وب‌کاوی، استخراج کلیدواژه، بازیابی اطلاعات، فرامتن

۳. کارشناس ارشد کتابداری و اطلاع‌رسانی و عضو هیئت علمی دانشگاه علوم پزشکی اصفهان
farshid_danesh@yahoo.com

1."Abstracting Keyword From Hypertext Docments".International Conference on Information and Knowledge Engineering, 2002, pp: 731-241.
2.Ben Choi & Baolin Li
4.HTML=Hypertext Markup Language

مقدمه

در مقالات علمی وتخصصی، کلیدواژه‌ها محتوای مقالات را مشخص می‌کنند. به‌طور کلی کلیدواژه‌ها مختصرترین شیوه خلاصه‌سازی محتوای مدارک است حتی به‌طور ناقص. خواننده به‌راحتی می‌تواند با بررسی موشکافانه کلیدواژه‌ها به حوزه‌های موضوعی تحت پوشش مقاله دست یابد. در این پژوهش پیشنهاد می‌شود که از کلیدواژه‌ها برای خلاصه‌سازی نتایج صفحات وبی که از موتورهای جستجو بازیابی می‌شوند، استفاده شود. هدف از این پیشنهاد کمک به کاربران برای بررسی سریع نتایج جستجو و بازیابی اطلاعات مورد نظر در زمان کوتاه‌تر است. اگرچه صفحات وب یا اسناد فرامتنی معمولاً فاقد سیاهه کلیدواژه‌ها هستند؛ این مقاله شیوه‌ای را به منظورجداسازی خودکارِ کلیدواژه‌های اسناد متنی یا فرامتنی ارائه می‌نماید.

این شیوه نه تنها همانند شیوه‌های دیگر، به بسامد واژه‌های موجود در مدرک توجه می‌کند بلکه، بسامد مترادف‌های واژه مورد نظر رانیز مد نظر قرار می‌دهد. شیوه پیشنهادی به عبارات کلیدی که شامل دو یا سه واژه هستند نیز توجه دارد. برای افزایش دقت در شمارش بسامد واژگان، الگوریتم ریشه‌یابی برای حذف پسوند واژه‌ها به‌کار برده شده است. همچنین این شیوه در مقایسه با پژوهش‌های مرتبط نسبتاً ساده است. پژوهشگران فنون پیچیده مختلفی را برای استخراج کلیدواژه‌ها از اسناد به‌کار برده‌اند مانند یادگیری ماشینی[5] (۱) ، نظـام فازی[6] (۲) ، شبکه‌های عصبی[7] (۳ ؛ ۴) و خود سازماندهی[8] (۵). برای مثال، زانگ[9] از روش شبکه‌های عصبی برای شناسایی خودکار کلیدواژه‌ها استفاده کرد. در این روش، شبکه‌ها برای شناسایی کلید واژه‌ها براساس روابط‌شان و یا واژه‌های هسته که به صورت دستی برای نشان دادن یک حوزه موضوعی انتخاب شده‌اند، آموزش داده می‌شوند. بعد از آموزش، شبکه‌های مورد نظر می‌توانند کلیدواژه‌ها را به‌طور خودکار از دیگر اسناد استخراج کنند و حوزه موضوعی مورد نظر را به مدارک اختصاص دهند (۳).

روش پیشنهادی، چندین مشخصه کلیدی را که در فنون آماری نسبتاً ساده یافت می‌شود در بردارد. شت[10] و دیگران روش‌های آماری زبان‌شناسی و روش‌های آماری وزن‌دهی به کلمات ارزشمند برای استخراج کلیدواژه‌ها را با هم مقایسه کردند. آنها دریافتند که کاربرد روش تحلیل آماری زبان‌شناسی بسیار پرهزینه است. بنابراین آنها استفاده از شیوه وزن‌دهی و شمردن ساده کلمات را برای استخراج کلیدواژه از مدرک پیشنهاد کردند (۶). هالت[11] و دیگران می‌گویند تحلیل بسامدِ اصطلاحاتی که در متن مدارک یافت می‌شود، ممکن است منبع اصلی دانش در مورد مدرک باشد. آنها همچنین پیشنهاد کردند اصطلاحنامه‌هایی که به صورت سلسله مراتبی سازماندهی شده‌اند درمقام دومین منبع دانش به حساب آیند (۷). جنکینز[12] و دیگران نیز تحلیل بسامد واژه‌ها را به‌کار بردند اما دریافتند وزن‌دهی به کلیدواژه به محلی که آن واژه در مدرک یافت می‌شود بستگی دارد (۸).

فرایند تولید کلیدواژه پیشنهادی بر مبنای فرضیات زیر بنا شده است:

- کلیدواژه، جزء واژگان غیرمجاز نباشد. واژه غیرمجاز واژه‌ای‌ست که معمولاً از آن استفاده می‌شود، اما زمانی که به صورت تنها استفاده می‌شود موضوع یا معنای خاصی ندارد.
- واژگان مختلفی که هم‌ریشه باشند، تقریباً دارای معانی و موضوعات یکسانی هستند. برای مثال، اتومبیل و اتومبیل‌ها مقوله یکسانی را شرح می‌دهند. روش ساده و مؤثراستخراج کلیدواژه می‌تواند بر پایه دگرگونی شکلی زبان باشد. رایج‌ترین شیوه ریخت‌شناسی ریشه‌یابی‌ست. فرایند ریشه‌یابی با از بین‌بردن پسوندها، ریشه‌ای که مفهوم واژه را تحت تأثیر قرار می‌دهد، حفظ می‌نماید.
- واژه‌هایی که بارها به‌کار برده می‌شوند موضوع مدرک را بسیار بیشتر از کلماتی که کمتر استفاده می‌شوند، بیان می‌کنند. معمولاً نویسنده برای توضیح موضوع مورد نظر خود از کلیدواژه‌هایی به‌طور مکرر استفاده می‌نماید.
- واژه‌ای که در متن زودتر پدیدار می‌شود، دارای اهمیت بیشتری در ارائه موضوع و محتوای سند است. معمولاً موضوع اصلی مدرک در قسمت اولیه آن ظاهر می‌شود.
- واژه و مترادف‌هایش، مفاهیم مرتبطی را ارائه می‌دهند برای خواندنی‌تر شدن مقاله، بیشتر نویسندگان مترادف‌های واژگان مختلف را برای جلوگیری از تکرار زیاد واژه‌ها به‌کار می‌برند.
- کلیدواژه می‌تواند مفرد یا ترکیبی از دو یا چند واژه مجاز

5.Machine Learning
6.Fuzzy system
7.Neural Networks
8.Self organization
9.Zhang
10.Sheth
11.Hulth
12.Jenkins

باشد البته ترکیب بیش از سه واژه برای ساختن کلیدواژه رایج نیست، بنابراین درفرایند تولید کلیدواژه، تعداد واژگان کلیدواژه یا عبارات بیش از سه کلمه نخواهد بود.

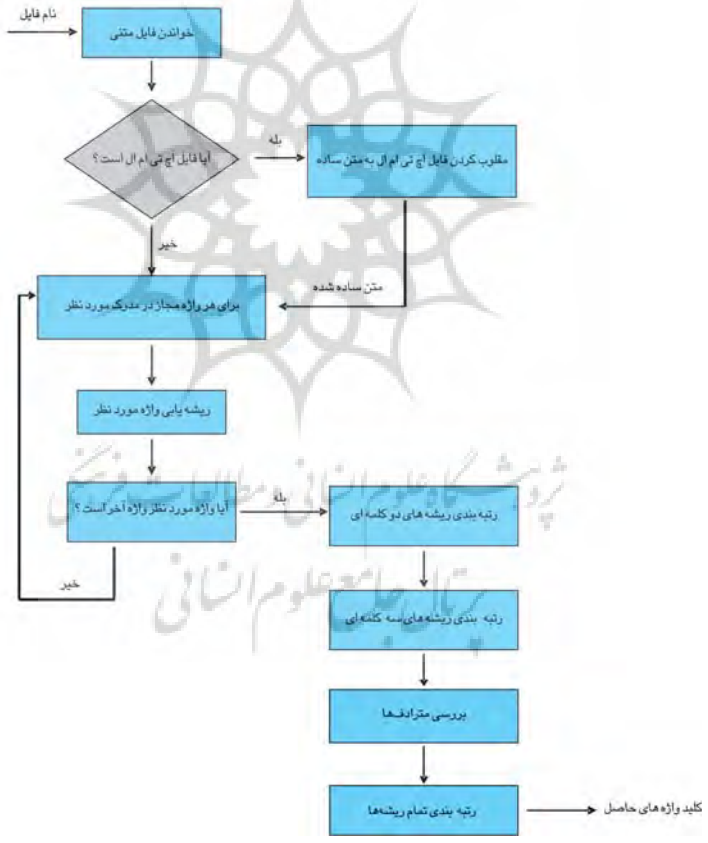
- کلیدواژه‌ای که به صورت واژه مفرد است، معمولاً از لحاظ دستوری قید و صفت نیست. صفت و قید به خودی خود کلیدواژه محسوب نمی‌شود. صفت‌ها و قیدها از لحاظ دستوری برای توصیف واژگان دیگر به‌کار می‌روند. با این وجود، زمانی که صفت با سایر واژگان ترکیب شود، می‌تواند کلیدواژه دو یا سه کلمه‌ای بسازد.

فرایند پیشنهادی در نمودار گردشی نشان داده می‌شود (تصویر ۱) و در ادامه نکات عمده آن مطرح می‌گردد.

فرایند پیشنهادی در نمودار چرخشی به صورت زیر است :

برای جداسازی کلیدواژه‌های مدارکی که به زبان اچ. تی. ام. ال هستند، ابتدا محتوای مدرک مورد نظر جدا می‌شود سپس برچسب‌ها و کدهای اچ. تی. ام. ال و واژه‌های غیرمجاز حذف می‌شود. سیاهه مترادف‌های واژه‌های باقی مانده مجاز تهیه می‌شود و پس از این مرحله، الگوریتم ریشه‌یابی برای از بین بردن پسوندها و باقی گذاردن واژگان مجاز به‌کار می‌رود. ریشه‌های باقی مانده براساس بسامد ریشه‌های موجود و مترادف‌های آنها گروه بندی می‌شوند.

در هنگام رتبه‌بندی ریشه‌ها، ریشه‌های تک‌واژه‌ای، ریشه‌های دوتایی و سه‌تایی متوالی برای ساختن سیاهه نهایی ریشه‌های کلیدواژه‌ها به کاربرده خواهد شد. این فرایند برای مدارک متنی نیز کاربرد دارد. در چنین حالتی، مرحله مقلوب‌سازی مدرکی که به زبان اچ. تی. ام. ال است و برای ساده‌تر کردن شکل متنی به‌کار می‌رود، نادیده گرفته خواهد شد و برنامه مستقیماً به مرحله ریشه‌یابی خواهد رفت (تصویر ۱). شیوه مورد نظر با جزئیات کامل در ادامه توضیح داده



شکل ۱. فرایند تولید کلید واژه ها

شده است:

ریشه یابی واژگان

دو نوع الگوریتم برای ریشه‌یابی واژگان وجود دارد که به‌طور گسترده، از آنها استفاده می‌شود: الگوریتم لووینز[13] و الگوریتم پورتر[14]. الگوریتم لووینز ۲۶۰ نمونه پسوند را مشخص می‌نماید و رویکردهای مکاشفه‌ای تکراری را به‌کار می‌برد.

الگوریتم پورتر ساده‌تر از الگوریتم لووینزاست. آن ۶۰ قاعده را که در گروه‌های خاصی سامان داده شده‌اند، به‌کار می‌برد. تضاد بین گروهی از قواعد، پیش از به‌کار بردن مجموعه قواعد دیگر برطرف می‌شود. همچنین قواعد مورد نظر در۵ مرحله متمایز، جدا می‌شوند و از مرحله ۱ تا ۵ شماره‌گذاری شده و برای واژگان مدارک به‌کار برده می‌شوند. درهر مرحله، گونه‌ای از پسوند واژه‌ها از بین خواهد رفت. پس از اتمام مرحله پنجم ریشه واژه‌ها باقی خواهد ماند. از آنجایی که الگوریتم پورتر ساده‌تر و سریع‌تر است، در این پژوهش الگوریتم مورد نظر به‌کار برده شد. همچنین طی این فرایند همه واژگان غیرمجاز و واژگانی که کمتر از سه حرف دارند، حذف خواهند شد.

رتبه‌بندی ریشه‌های تک کلمه‌ای

مجموع بسامدها برای رتبه‌بندی ریشه‌های تک کلمه‌ای استفاده خواهد شد. از ضریبِ محل، برای بالا بردن امتیاز ریشه‌هایی که در قسمت اولیه مدرک مورد نظرقرار گرفته‌اند، استفاده می‌شود. درصورتی‌که ریشه در یک سوم ابتدایی مدرک ظاهر شود، ضریب ۱/۳ را به خود اختصاص خواهد داد؛ اگر ریشه ای برای اولین بار در قسمت میانی مدرک پدیدار شود، ضریب ۱ برای ریشه مورد نظر در نظر گرفته خواهد شد؛ و اگر ریشه در یک سوم انتهایی مدرک مشاهده شود، ضریب ۰/۸ برای آن ریشه تعیین خواهد شد. از آنجا که کلیدواژه‌های تک‌واژه‌ای به صورت قید یا صفت به کار نمی‌روند، در نتیجه ضریب آنها صفر خواهد بود. مجموع امتیاز ریشه برابر است با مجموع بسامدها ضربدر ضریب محل. ریشه‌ها بر مبنای امتیاز نهایی‌شان رتبه‌بندی خواهند شد.

رتبه‌بندی ریشه‌های دو یا سه کلمه‌ای

پس از به‌دست آمدن ریشه‌های تک کلمه‌ای، ریشه‌های عبارات دو یا سه کلمه‌ای بررسی می‌شود. برای ریشه‌های دو کلمه‌ای، اگرکلمه اول صفت باشد، عبارت مورد نظرکلیدواژه به‌شمار می‌آید؛ اما اگر کلمه دوم عبارت دوکلمه‌ای صفت یا قید باشد و همچنین اگر دومین یا سومین کلمه عبارت سه کلمه‌ای، صفت باشد، امتیاز عبارات مورد نظر صفرخواهد بود و کلیدواژه محسوب نمی‌شود.

بررسی مترادف‌ها

روش پیشنهادی این پژوهش، مترادف‌های هر یک از واژگان مجاز را بررسی می‌کند. همچنین با توجه به تعداد مترادف‌هایی که در مدرک تحت بررسی، برای واژه مورد نظر وجود دارد رتبه واژه مزبور افزایش می‌یابد. از آنجا که مترادف یک واژه اهمیت کمتری نسبت به خود واژه دارد، ضریبی که به آن اختصاص داده می‌شود کمتر از ۱ (یک) است. در این پژوهش ضریب ۰/۸ برای هر یک از مترادف‌ها در نظر گرفته شده و مترادف‌های مورد استفاده از سایت http://www.thesarus.com به‌دست آمده است (۹).

رتبه‌بندی همه ریشه‌ها

همه ریشه‌هایی که شامل ریشه‌های تک کلمه‌ای، ریشه‌های دو یا سه کلمه‌ای هستند برطبق مجموع امتیازاتشان مرتب خواهند شد. کلمات یا عباراتی که دارای بالاترین امتیاز باشند، کلیدواژه‌های مدرک محسوب می‌شوند. در این شیوه، کاربران می‌توانند تعداد کلیدواژه‌های حاصل را مشخص نمایند؛ در غیر این‌صورت تعدادی از کلیدواژه‌ها به صورت پیش‌فرض فراهم خواهد شد.

نتایج آزمون‌ها

فرایند پیشنهادی بررسی و آزمایش شد. در این بررسی از ۲۰ مدرک متنی و اچ . تی. ام. ال که دامنه لغات آنها بین صد تا هزاران واژه بود، در آزمون استفاده شد. نتایج آزمون نشان داد که میان زمان محاسبه شده برای انجام فرایند مورد نظر و

13.Lovins
14.Porter

تعداد واژگان موجود در مدارک، نسبت خطی وجود دارد (تصویر۲). همچنین آزمایش نشان داد که ۵۶/۷ درصد از کل زمان، صرف ریشه‌یابی واژگان می‌شود.

برای تعیین درستی و دقت این شیوه از مقالات علمی وتخصصی که دارای سیاهه کلیدواژه‌ها بودند؛ استفاده شد (۱۰). نتایج آزمایش نشان داد که شیوه مورد نظر به‌طور متوسط می‌تواند ۵۲ درصد از کلیدواژه‌هایی را که در مقالات فهرست شده‌اند، فراهم نماید.

نتیجه‌گیری

در این مقاله شیوه جداسازی‌های مدارک متنی و فرامتنی توضیح داده شد. این شیوه مترادف‌ها و موقعیت واژگان را در مدارک به حساب می‌آورد. پیشنهاد می‌شود که از کلیدواژه‌ها، به منظور خلاصه‌سازی نتایج صفحات وبی که به‌وسیله موتورهای جستجو بازیابی شده‌اند، استفاده شود. هدف از این پیشنهاد یاری رساندن به کاربران برای بررسی سریع نتایج جستجو و یافتن اطلاعات مورد نظر در مدت زمان کوتاهتر است.

در نظام مورد نظر برای پردازش تعداد زیادی از صفحات وبی که از موتورهای جستجو بازیابی شده‌اند و فراهم کردن زمان مناسبی برای پاسخ، شیوه نسبتاً ساده‌ای برای این تحقیق انتخاب گردید. با این وجود نتایج آزمایش نشان داد حتی چنین شیوه ساده‌ای هم، از نظر زمان پردازش و دقت در انجام فرایند، نیازمند اصلاحات بیشتری‌ست. بنابراین ضروری‌ست پس از جستجوی انجام شده به جای پیش‌پردازش به شیوه تولید کلیدواژه‌ها، پیش‌پردازش به شیوه جداسازی و ذخیره کلیدواژه‌ها در پایگاه داده‌ها صورت گیرد. این مقاله بخشی از پژوهش در حوزه وب‌کاوی‌ست، البته زمینه‌های پژوهشی زیادی در این حوزه وجود دارد که پرداختن به آنها در آینده، اجتناب‌ناپذیر است.

منابع

1. Azcarraga, A.; Yap Teddy, J. "Comparing keywords extraction techniques for WEBSOM text archives". In The 13th IEEE international conference on tools with artifcial intelligence, Dallas, (USA, 7-9 November 2001).

2. Chau, R.; Yeh, Ch. "Explorative multilingual text retrieval based on fuzzy multilingual keyword classification". In The 5th international workshop on information retrieval with Asian languages, (November 2000).

3. Chung, Y; Pottenger, W.; Schatz, B. "Automatic subject indexing using an associative neural network". In The 3th ACM conference on digital libraries, (May 1998).

4. Hulth, A. … [et al]. "Atomatic keyword extraction using domain knowledge". In The 2nd international conference

تصویر ۲. ارتباط بین زمان پردازش و تعداد واژگان

7. SC96. Technical paper abstracts. [on-line]. Available:
www.supercomp.org/sc96/proceedings/sc96proc/tabst.htm

8. Sheth, S.; Yau, D. "Smart scope: Intelligent keyword generation". 2002. [on-line]. Available: http://sydewww.uwaterloo.ca/ underGrad/workshop/1999-2000/smartscope.html

9. Thesarus.com. [on-line]. Available: http://www.thesaurus.com

10. Zhang, Sh.; Powell, H.; Palmer-Brown, D. "Keyword extraction using an artificial neural network". 2000. [on-line]. Available: www.uilots.let.un.nl/~paola.zhang.html

CLCLing, Mexico City, (Mexico, 18-24 Febuary 2001).

5. Jenkins, C. … [et al]. "Atomatic RDF metadata generation for resource discovery". In The 8th international WWW conference, Toronto, Canada. 1999. [on-line]. Available: http://scit.wlv.ac.uk/~ex 1253/rdf_ paper.

6. Maldenic, D.; Grobelink, M. "Assigning keywords to documents using machine learning". In The 10th international coference on information anf intelligent system ISS-99, Varazdin, Crotia, (September 1999).