به سوی نمایه‌سازی مفهومی با استفاده از تخصیص خودکار توصیفگرها[1]

آرتر مونژو رایز[2]
ترجمه حسن اشرفی ریزی[3]

چکیده

فنون نمایه‌سازی به تکامل خوبی رسیده است. کتابخانه‌ها و دیگر مجموعه‌های رقومی برای ذخیره و بازیابی مدارک از این فنون بهره فراوان می‌برند. از طرف دیگر ما فنونی داریم که به کاربر اجازه گردآوری اطلاعات را می‌دهد. رویکردهای جاری به اندازه کافی برای رضایت کاربر پیشرفته نیستند. در "آزمایشگاه فیزیک ذرات اروپا[4] ما مشغول ایجاد یک برنامه نمایه‌ساز مبتنی بر توصیفگرهای اصطلاحنامه هستیم. با مجموعه‌ای از مدارک مرتبط با اصطلاحنامه، کاربر می‌تواند آنها را به شیوه مفهومی‌تر بازیابی کند. در این مقاله به توصیف هسته این نظام، یعنی تخصیص‌دهنده توصیفگر خودکار پرداخته می‌شود.

کلیدواژه‌ها

نمایه‌سازی مفهومی، نمایه‌ساز فیزیک با انرژی بالا[5]، جستجوی تمام متن، تخصیص خودکار توصیفگرها

مقدمه

فنون نمایه‌سازی توجه محققان بازیابی اطلاعات را جلب کرده است، زیرا به‌وضوح بیان‌کننده یکی از مشکلات اصلی‌ست که باید رفع، و بهینه‌سازی شود. درک چنین توجهی آسان است، زیرا نمایه‌سازی تأثیر زیادی بر بقیه اجزای نظام بازیابی اطلاعات دارد. با استفاده از تجربیات پیشین ریجسبرگن[6] و سالتون[7] در زمینه بازیابی اطلاعات، می‌توان فرایند دسترسی به اطلاعات را به‌صورت زیر خلاصه کرد:

۳. کارشناس ارشد کتابداری و اطلاع‌رسانی مرکز اطلاع‌رسانی و داده‌پردازی سازمان انرژی اتمی ایران
hassanashrafi81@yahoo.com

1." Towrd conceptual indexing using au-tomatic assignment of descriptors "/[on -line] Avinable: http://www.dimi.Uniud.it/- mizzaro/AH2002/proceedings/pdfs/Bmontejo .PDF

2. Artur Montejo Raez
4.European Organization for Nuclear Research
5 .High Energy Physics Indexer
6.Rijsbergen
7.Salton

۱. کاربرنیاز اطلاعاتی خاصی دارد.
۲. کاربر نیاز اطلاعاتی خود را به نظام منتقل می‌کند.
۳. نظام به مجموعه دسترسی دارد و مجموعه‌ای از مدارک را برای کاربر بازیابی می‌کند.
۴. کاربر مجموعه را مرور می‌کند و بازخورد را به نظام برمی‌گرداند.
۵. نظام، بازخورد را از کاربر دریافت می‌کند، بنابراین می‌تواند جستجوی بهتری انجام دهد.
۶. گفتگوی نظام و کاربر زمانی تمام به‌پایان می‌رسد که کاربر از نتایج به دست آمده راضی باشد (۱۹؛ ۱۷).

جستجوی تمام‌متن با عنوان "هسته فلسفی"[8] برای این گفتگو در نظر گرفته شده است. چندین زبان پرسشی می‌توان یافت که این جستجو را تقویت کند تا کاربر پرسش‌های دقیق‌تری مطرح سازد. مشکل اصلی ابهام پرسش است، بدین معنا که کاربر هنگام مشخص‌کردن نیازهای اطلاعاتی خود هیچ کوششی انجام نمی‌دهد. برخی رویکردها، برای رفع این مشکل، برای تهیه رابط بهتری برای جستجو اقدام کرده‌اند و به نظر می‌رسد که ترکیبی مناسب از الگوریتم‌های رتبه‌بندی مانند رتبه‌بندی صفحه‌ای[9] (۱۳) همراه با ابزارهای مستقیم و سریع مرور، مانند دسته‌بندی مجموعه نتایج مناسب‌ترین راه حل باشد.

جستجوی مدارک در محیط‌های پژوهشی، مانند "آزمایشگاه فیزیک ذرات اروپا"، کار پیچیده‌ای‌ست، زیرا لازمه آن گردآوری مجموعه عظیمی‌ست. اکنون بیشتر روی توسعه ابزارهای معناشناختی کار می‌کنیم تا برای محققان توانایی حرکت از یک مدرک به مدرک دیگر بر مبنای مفهوم مهیا شود. ما از اصطلاحنامه دسی[10] برای انتخاب توصیفگرهای مرتبط با مدارک "فیزیک با انرژی بالا" استفاده می‌کنیم. در این شیوه ابرداده‌هایی اضافه می‌شوند که اطلاعاتی مرتبط با محتوای معناشناختی مدرک ارائه دهند. از آنجا که همه توصیفگرها متعلق به اصطلاحنامه ساختاریافته هستند، بنابراین مدارک یکدست می‌شوند (۶). حال می‌توان به شبکه‌ای با هدف وب معنایی که پیشنهاد تیم برنزرلی[11] و همکارانش بود، به کتابخانه رقومی در این حوزه اندیشید (۳).

۲. اضافه کردن ابرداده‌های معناشناختی به مدرک توصیفگرها

مقالات حاوی اطلاعات در موضوعی خاص هستند که توسط نویسندگان عرضه شده‌اند(معمولاً فقط زمانی که نسبت به انتشار آنها در مجله اقدام می‌شود). برخی نشریات کلیدواژه دارند و تعداد اندکی هم رده‌بندی «طرح رده‌بندی فیزیک و ستاره‌شناسی»[12] را که مورد حمایت جامعه فیزیک آمریکا[13]ست، پذیرفته‌اند (۱)؛ اما بعید است که از این رویکردها در همه مدارک به‌طور کامل استفاده شود، بنابراین برای جستجو در مقیاس جهانی مفید نیستند. به همین دلیل هر نوع داده اضافه شده، باید توسط پدیدآورندگان آن پایگاه اطلاعاتی برای جستجوی کارآمد تهیه شود.

این شیوه اضافه‌کردن منابع موضوعی، "نمایه‌سازی موضوعی"یا "تقویت کلیدواژه‌ها"[14] نامیده می‌شود. دو شیوه متفاوت برای این کار وجود دارد: واژه‌ها را از یک اصطلاحنامه مشخص انتخاب کرد، یا کلیدواژه‌هایی را به کار برد که بنا به تمایل نمایه‌ساز می‌توانند انتخاب شوند. اختصاص کلیدواژه‌های مناسب از یک اصطلاحنامه مشخص، نیاز به نمایه‌ساز را تشدید می‌کند، زیرا مفهوم مدارک باید به‌خوبی درک شود. ممکن است واژه‌های به‌کار رفته در نمایه اصلاً در متن ظاهر نشوند، که این روش می‌تواند فایده بیشتری نسبت به هر راهبردی که فقط از کلمات متن مدرک استفاده می‌کند، داشته باشد. نمونه اصطلاحنامه‌های معتبر آنهایی هستند که توسط نظام بین‌المللی اطلاعات هسته‌ای[15] (۹) و چکیده‌نامه فیزیک، کامپیوتر، و مهندسی الکترونیک INSPEC و استفاده می‌شوند (۱۰)، و اصطلاحنامه دسی هم از این دسته اصطلاحنامه‌هاست.

مدارک جدید زیادی هر روز به کتابخانه سرن[16] می‌رسد که تقریبا همگی الکترونیکی هستند. نمایه‌سازی به وسیله نمایه‌سازانی که در دسی کار می‌کنند، انجام می‌شود. به دلیل افزایش مقالات مربوط به HEP ، رویکردی جدید برای تخصیص کلیدواژه‌ها به وجود آمده است. چون هنوز نمایه‌سازان خودکار راه زیادی تا راه‌حل نهایی دارند، می‌توان از ابزارهای کمکی مبتنی بر رایانه برای نمایه‌سازی به منظور تسهیل کار نمایه‌سازان انسانی استفاده کرد.

برنامه "نمایه‌ساز فیزیک با انرژی بالا" قصد دارد که راه حلی

8.Philosophal stone
9.Page Rank
10.DESY
11.Tim Berners Lee
12.Physics and Astronomy Classification Scheme
13.American Physical Society
14.Keyword enhancement
15.INIS = International Nuclear Information System
16.CERN

مقدماتی پیشنهاد کند و راه را برای پژوهش در زمینه ابزارهای نمایه‌سازی خودکار در حوزه "فیزیک با انرژی بالا" باز کند. این برنامه توصیفگرهایی را برای مدارک مشخصی پیشنهاد می‌کند. اولین گام در توسعه چنین نظامی یعنی تولید کلیدواژه‌های اصلی دسی برداشته شده است. این توصیفگرها بر مبنای رویکردهای آماری به وجود آمده‌اند (۱۹).

*coherent interaction
state (for quantum mechanical states)
coherent
cohomology
* coil
- coincidence('fast logic'or 'trigger' or
'associated production')
- Coleman - Glashow formula
(baryon , mass difference)
- coleman - Weinberg instabillity
(symmetr breaking)
* collective
(used only in connection with accelerators)
* Collective phenomena ('field theory,
collective phenomena ' or 'nuclear physics,
collective phenomena'or 'nuclear matter,
collective phenomena')
- collider ('storage ring' or 'linear collider')
Colliding beam detector
(use only in instrumental papers)
* colliding beam (for accelerator use
'storage ring ' or color(for colored partons)
colored particle
communications

شکل ۱. اقتباسی از تزاروس دسی

شکل ۱ اقتباسی از تزاروس دسی‌ست. توصیفگرهایی که با ستاره مشخص شده‌اند، کلیدواژه‌های توصیفی و ثانوی هستند؛ آنهایی که با خط تیره آمده‌اند کلیدواژه نیستند، و هر چه قبل از آن جای خالی وجود دارد کلیدواژه اصلی‌ست.

۳.پیشینه کار

امکان دسترسی به مجموعه عظیمی از مدارک به‌صورت تمام‌متن نشان‌دهنده ظهور عصری جدید در بازیابی اطلاعات است. بیشتر پژوهش‌ها درباره پردازش زبان طبیعی‌ست. اولین اثر سالتون (۱۶) مقدمه جالبی در این زمینه ارائه می‌دهد. بسیاری از الگوریتم‌ها، از رویکرد "الگوریتم‌های مرتبط" ناشی شد؛ از الگوریتم‌های ترکیبی کلاسیک[17] برای کاهش عرضه یک مدرک به مقوله‌های اساسی آن (به منبع ۱۵ نگاه کنید)؛ تا آنهایی که با مدرک به‌مثابه یک کل، برای تشخیص قالب کلام (۱۲) یا عبارت‌های مفهومی رفتار می‌کنند (۵).

درباره تخصیص توصیفگرها، دو رویکرد متفاوت وجود دارد: آنهایی که با هدف به‌کار گرفته شدن توسط انسان ارائه شده‌اند وآنهایی که به وسیله دیگر ابزار محاسباتی به‌کار گرفته می‌شوند. برای اولین رویکرد می‌توان به نظام‌هایی مانند مش[18]، بایوسیس[19]، ناسا[20]، مای[21] که در گذشته توسعه یافته‌اند، اشاره کرد (۸). دومین رویکرد، رویکردهایی چون رویکرد احتمالگرایی رجینالد فربر[22] (۷) و برخی رویکردهای چندزبانه، مانند "نمایه‌ساز استفاده‌شده در کمیسیون اروپایی"[23] (۱۸) برای اهداف بین‌زبانی و نظام ماژیک[24] متعلق به کاتشمانش[25] و دیگران را دربرمی‌گیرد (۱۱).

بهترین شیوه استفاده از توصیفگرها، ترکیبی از این دو رویکرد است. این دو رویکرد به ما امکان ایجاد ارتباط درونی بین مدارک، و به کاربران اجازه گردآوری و استفاده از آنها را می‌دهند. نظام "نمایه‌ساز فیزیک با انرژی بالا"، هسته همه این موارد است و به صورت خودکار توصیفگرهایی برای مدارک تمام متن ارائه شده، پیشنهاد می‌کند.

۴.نمایه‌ساز "فیزیک با انرژی بالا"

الگوریتم به کار رفته به مجموعه‌ای از داده‌ها نیاز دارد که باید از قبل در فرایند آموزش تولید شده باشد. پس از آن، نظام قادر خواهد بود تا توصیفگرهای اصلی را با درجه قابل قبولی از موفقیت - که از طریق آزمون ثابت شده- پیشنهاد کند. این دو مرحله به مجموعه‌ای از مدارک در مقام درونداد نیاز دارند. نمایه‌ساز "فیزیک با انرژی بالا" با مجموعه‌ای بالغ بر ۳۷۰۰ مدرک تمرینی تأمین شده است. این مجموعه نمونه‌ای

17.Classic conflation algorithms
18.MeSH
19.Biosis
20.NASA
21.MAI
22.Reginald Ferber
23.European Commission
24.MAGIC
25.Kutschekmanesch

از مدارک مرتبط با "فیزیک با انرژی بالا" ست که در آن توصیفگرهای دسی به هر مدرک اختصاص داده شده است. یعنی فهرستی از مدارکی داریم که قبلا به وسیله توصیفگرهای دسی برچسب خورده است و نظام ما می‌تواند از آن استفاده کند. بعد از آموزش می‌توان مدارک جدیدی به نظام وارد کرد و برونداد آن را به صورت فهرستی از توصیفگرهای پیشنهادی خودکار دریافت داشت.

۴-۱. الگوریتم

آموزش شامل موارد زیر است:

۱. هر مدرکی تجزیه و ترکیب می‌شود، کلمات غیرمجاز[۲۶] آن (حروف تعریف، حروف اضافه، و دیگر کلمات بدون معنی) حذف می‌شود، و ریشه یاب[۲۷] (برای به‌دست آوردن ریشه هر کلمه) به کار برده می‌شود. بالاخره بسامد هر واژه باقی مانده در مدرک محاسبه می‌شود.

۲. برای هر توصیفگر، برداری از اصطلاحات را با استفاده از فرمول زیر محاسبه می‌کنیم:

$$\text{Weight}(k,t) = \lg\frac{M}{M_t}\sum \text{TFJD}_{t,d}^{KF_K}$$

- Weight (k, t) وزن اصطلاح t برای توصیفگر k است.
- M مجموع تعدادکل توصیفگرهاست.
- M_t تعداد توصیفگرهای مرتبط باواژه t است (یعنی اصطلاح t در مدارک Mt ظاهر می‌شود، مدارکی که برچسب k دارد.
- d یک مدرک است.
- $TFIDF_{t,d}$ بسامد مدرک ضربدر عکس بسامد مدرک اصطلاح t در مدرک d است.
- $KF_{k,}$ میزان توصیفگر K در مدرک d است.

با تخصیص توصیفگرهای دا ده شده به مدرک جدید ،همه توصیفگرهای موجود در اصلاحنامه با وزن محاسبه‌شده در زیر رتبه‌بندی می‌شوند:

۱. مدرک مانند مرحله آموزش، تجزیه و ترکیب می‌شود، تا برداری از اصطلاحات بر اساس بسامد به دست آید.

۲. این بردار در ماتریسی از وزن‌های موجود میان توصیفگرها و واژه‌ها ضرب می‌شود، در نتیجه برداری از اصطلاحات وزن داده شده به دست می‌آید.

۴-۲. نتایج

این نظام از طریق رابط‌های مبتنی بر وب[۲۸] با کاربر تعامل دارد. کاربر با استفاده از مرورگرهای وب می‌تواند نظام را با مدارکی از مجموعه آزمایشی امتحان کند یا کلیدواژه‌های پیشنهادی را با ارئه مدرک جدید تمام‌متن، در قالب پست‌اسکریپ[۲۹] یا پی. دی. اف[۳۰] یا متن معمولی به دست آورد. اگر چه این نظام فقط می‌تواند توصیفگرهای اصلی دسی را پیشنهاد دهد، اما نتیجه از نظر دقت و بازیابی در حد۶۰ درصد است. یعنی به‌طور میانگین ۶۰ درصد از کلیدواژه‌های پیشنهادی همان‌هایی هستند که در فهرست ارائه شده توسط دسی، یافت می‌شود و ۶۰ درصد کلیدواژه‌های ارائه‌شده توسط دسی همان‌هایی هستند که توسط نظام ارائه می‌شوند.

این نظام هم اکنون در خدمت‌دهنده[۳۱] مدارک "آزمایشگاه فیزیک ذرات اروپا" ادغام شده است (۴). از آنجا که برنامه در مراحل ابتدایی قرار دارد، اصلاحاتی باید صورت گیرد. کلیدواژه‌های ثانویه و الگوریتم‌های پالایش شده[۳۲] (با استفاده از منابع زبان‌شناسی) به‌منظور افزایش عملکرد این نظام در دست مطالعه و بررسی هستند.

نتایج و کارهای آینده

هنوز باید پیشرفت‌هایی در این نظام انجام گیرد. برنامه نمایه‌سازی جدید "نمایه‌ساز فیزیک با انرژی بالا" اکنون مبتنی بر جاوا و مای‌اسکوئل[۳۳] برنامه‌ریزی می‌شود. مقیاس‌های دیگر آزمایش می‌شوند و تصمیم گرفته شده تا قابلیت پرداختن به چند کلمه‌ای‌ها یکدست شود. سپس، نظام آماده گنجیدن در ابزار جستجو خواهد بود و مشخصه تجمیع مدارک و استنادها را با توصیفگرهای اصلاحنامه به‌مثابه ارزش افزوده فراهم می‌کند.

منابع

1. American Institute of Physics. "Physics and astronomy classification scheme".

26.Stop Words
27.Stemmer
28.Web-based interface
29.Postscript
30.PDF
31.Server
32.Refined algorithms
33.MYSQL

http://www.iaea.or.at/worldatom/ publications/inis/inis.html

10. "Inspec thesaraus". [on-line]. Available: http://www.iee.org.uk/publish/inspect/ .

11. Kutschemanesch, Said … [et al]. "Automated multilingual indexing: A synthesis of rule-based and thesaurus-based methods". In Pub Deutschen Gesellschaft fur Dokumentation, editor, ***Information und Markte***, pp: 211-224, Donn, Germany, 1998.

12. Marcu, Daniel. "Discourse trees are good indicators of importance in text". Technical report, Information Science Institute, University of Southern California, 1997.

13. Page, Lawrence … [et al]. "The page rank citation ranking: Bringing order to the web". Technical report, Computer Science Department, Stanford University, 1998.

14. Palmer, Christopher ... [et al]. "Demonstration of hierarchical document clustering of digital library retrieval results". ***In ACM/IEEE Joint Conference on Digital Libraries***, 2001, p. 451.

15. Robertson, A. M.; Willett, P. "Evaluation of techniques for the conation of modern and seventeenth century English spelling". In Tony McEnery and Chris Paice, editors, ***Proceedings of the BCS 14th Information Retrieval Colloquium***, Workshops in Computing, pp: 155-168, London, April 13-14 1993.

16. Salton, G. "Automatic text analysis". Technical Report TR69-36, Cornell University [on-line]. Available: http://publish.aps.org/ PACS/.

2. Baeza-Yates, R.; Riberio-Neto, Berthier. ***Modern Information Retrieval***. [s. n.]: Acm Press Series,1999.

3. Berners-Lee, Tim; Hendler, James; Lassila, Ora. "The semantic Web". ***Scientific American***, Vol. 284, No. 5 (May 2001):34-43.

4. CERN. DH Group, ETT division. "The cern document server". 1996. [on-line]. Available: http://cds.cern.ch.

5. Culy, Christopher. "An extension of phrase structure rules and its application to natural language". Master's thesis, Stanford University, 1983.

6. DESY. "The high energy physics index keyword". 1996. [on-line]. Available: http://www.library.desy.de/schlagw2.html.

7. Ferber, Reginald. "Automated indexing with thesaurus descriptors: a cooccurrence-based approach to multilingual retrieval". In Carol Peters and Costantino Thanos, editors, ***Proceedings of ECDL-97, 1st European Conference on Research and Advanced Technology for Digital Libraries***, pp: 233-251, Pisa, IT, 1997. Lecture Notes in Computer Science, number 1324, Springer Verlag, Heidelberg, DE.

8. Gail Hodge, Oak Ridge. "Cendi agency indexing system descriptions: A baseline report". Technical report, CENDI. 1998. [on-line]. Available: http://www.dtic.mil/ cendi/publications/98-2index.html.

9. "Inis thesaraus". [on-line]. Available:

Natural Language Processing (SEPLN'2001), Jan (Spain), September 2001, pp: 273-280.

19. Van Rijsbergen, C. J. ***Information Retrieval***. London: Butterworths, 1975. [on-line]. Available: http://www.dcs.gla.ac.uk/Keith/Preface.html.

sity, Computer Science Department, June 1969.

17. Ibid. "A vector space model for automatic indexing", 1975.

18. Steinberger, Ralf. "Cross-lingual keyword assignment". In L. Alfonso Ure na Lopez, editor, ***Proceedings of the XVII Conference of the Spanish Society for***