خوشه‌بندی سلسله‌مراتبی فازی برای کشف روابط معنایی پنهان در اسناد وب‌معنایی

بهنام طاهری خامنه و حمید شکرزاده*

گروه مهندسی کامپیوتر، واحد پردیس، دانشگاه آزاد اسلامی، پردیس، ایران

چکیده

رشد انبوه اطلاعات در وب مشکلاتی را به‌دنبال داشته است که از مهم‌ترین آن‌ها می‌توان به چالش‌های ایجادشده برای جستجو در وب اشاره کرد. با توجه به این که بیشتر محتویات وب امروزی برای استفاده توسط انسان طراحی شده است، ماشین‌ها تنها قادر به دست‌کاری و فهم داده‌ها در سطح لغت هستند؛ این مسأله مهم‌ترین مانع در سرویس‌دهی بهتر به کاربران وب است. هدف این مقاله ارائه‌ی نتایج بهتر در پاسخ به جستجوی کاربران وب معنایی است. به این منظور در روش پیشنهادی ابتدا عبارت مورد نظر کاربر با توجه به میزان موضوعات مرتبط با آن، مورد بررسی قرار می‌گیرد. پاسخ به‌دست‌آمده از این بررسی، وارد یک سامانه رتبه‌دهی متشکل از سامانه تصمیم‌گیری فازی و خوشه‌بندی سلسله‌مراتبی می‌شود تا نتایج مطلوب‌تری را به کاربر بازگرداند. گفتنی است که روش پیشنهادی نیاز به هیچ‌گونه دانش قبلی برای خوشه‌بندی داده‌ها ندارد؛ علاوه‌بر این دقت و جامعیت این پاسخ نیز اندازه‌گیری می‌شود؛ درنهایت، بر روی نتایج به‌دست‌آمده آزمون F اعمال می‌شود که اغلب به‌عنوان یک معیار از عملکرد سامانه، برای ارزیابی الگوریتم و سامانه‌های مورد استفاده در نظر گرفته می‌شود. نتایج حاصل از این آزمون نشان می‌دهد که روش ارائه‌شده در این مقاله می‌تواند پاسخ دقیق‌تر و جامع‌تری نسبت به روش‌های مشابه خود ارائه دهد و به‌طور میانگین دقت را تا ۱/۲۲ درصد افزایش دهد.

واژگان کلیدی: وب معنایی، منطق فازی، خوشه‌بندی سلسله‌مراتبی، روابط معنایی پنهان، الگوریتم HFCS

Hierarchical Fuzzy Clustering Semantics (HFCS) in Web Document for Discovering Latent Semantics

Behnam Taheri Khameneh & Hamid Shokrzadeh*

Department of Computer Enginiering, Pardis Branch, Islamic Azad University, Pardis, Iran

Abstract

This paper discusses about the future of the World Wide Web development, called Semantic Web. Undoubtedly, Web service is one of the most important services on the Internet, which has had the greatest impact on the generalization of the Internet in human societies. Internet penetration has been an effective factor in growth of the volume of information on the Web. The massive growth of information on the Web has led to some problems, the most important one is search query. Nowadays, search engines use different techniques to deliver high quality results, but we still see that search results are not ideal. It should also be noted that information retrieval techniques to a certain extent can increase the search accuracy. Most of the web content is designed for human usage and machines are only able to understand and manipulate data at word level. This is the major limitation for providing better services to web users. The solution provided for this topic is to display the content of the web in such a way that it can be readily understood and comprehensible to the machine. This solution, which will lead to a huge transformation on the Web is called the Semantic Web and will begin. Better results for responding to the search for semantic web users, is the

* Corresponding author

*نویسندهٔ عهده‌دار مکاتبات

• تاریخ ارسال مقاله: ۱۳۹۷/۴/۲۳ • تاریخ پذیرش: ۱۳۹۸/۴/۱۹ • تاریخ انتشار: ۱۳۹۹/۰۴/۰۱ • نوع مطالعه: کاربردی

[DOI: 10.29252/jsdp.17.1.29]

purpose of this research. In the proposed method, the expression, searched by the user, will be examined according to the related topics. The response obtained from this section enters to a rating system, which is consisted of a fuzzy decision-making system and a hierarchical clustering system, to return better results to the user. It should be noted that the proposed method does not require any prior knowledge for clustering the data. In addition, accuracy and comprehensiveness of the response are measured. Finally, the F test is applied to obtain a criterion for evaluating the performance of the algorithm and systems. The results of the test show that the method presented in this paper can provide a more precise and comprehensive response than its similar methods and it increases the accuracy up to 1.22%, on average.

Keywords: Semantic Web, Fuzzy Logic, Hierarchical Clustering, Latent Semantic, HFCS

۱- مقدمه

موتورهای جستجو، ابزاری ضروری برای یافتن و استخراج اطلاعات مورد نظر از صفحات وب است که در جمع‌آوری و پالایش[1] اطلاعات به کاربران کمک می‌کند [1]. در طول دهه گذشته، بهینه‌سازی بی‌وقفه در روش‌های بازیابی اطلاعات، موتورهای جستجو را به سطوح جدیدی از کیفیت رهنمون کرده است؛ به‌طوری‌که جستجوی وب به یک منبع استاندارد و اغلب ارجح برای یافتن اطلاعات تبدیل شده است [2]. اسناد وب در حالت عادی ناهمگن و پیچیده است و ارتباط پیچیده‌ای میان اجزای یک سند و اجزای یک سند با سندی دیگر، وجود دارد. فعل و انفعالات بالا بین اصطلاحات در اسناد گاهی نشان‌گر معانی مبهم است. این ابهام نه‌تنها برای کاربر بلکه برای موتورهای جست‌وجو نیز وجود دارد [3]. وب حاضر مجموعه‌ای از مستندات است که با پیوندهایی به هم مرتبط شده است و هیچ معنی و مفهومی بین این ارتباطات حاکم نیست. در این وب تا زمانی که پیوندها انتخاب نشود، هیچ نظری در مورد محتوای آن نمی‌توان اظهار کرد. ساختار نمایشی وب به‌گونه‌ای طراحی شده است که نمی‌تواند محتوا را برای ماشین متمایز کند به همین دلیل است که موتورهای جستجو امکان درک آن را ندارند که یک صفحه وب درباره چه موضوعی صحبت می‌کند. درواقع آن‌ها فقط تعدادی کلیدواژه و عبارت را نمایه[2] می‌کنند و پس از آنکه کاربر، پرس‌وجوهایی را ارسال کرد، بر اساس آن‌ها نتایجی را ارائه می‌دهند. کاربران وب بیشتر علاقه‌مند به کشف مفاهیم اطلاعات در وب هستند تا استفادهٔ مستقیم از آن‌ها، درواقع کاربران از موتورهای جستجو توقع درک متن موردنظر خود را دارند؛ بنابراین ما به‌دنبال بازنمایی از اطلاعات هستیم که در آن ماشین با سرزدن به یک تارنما معنای هر یک از بخش‌های[3] موجود در آن را درک کند. وب‌معنایی به‌دنبال راهی می‌گردد تا کاربر پرس‌وجوی مورد نیازش را به مرورگر بدهد و رایانه در پشت پرده به جستجو و ارسال و دریافت اطلاعات متفاوتی بپردازد و خروجی را در شکلی جدید به کاربر اعلام کند تا کاربر دیگر نیازی به جستجو در نتایج یافت‌شده نداشته باشد. این امر نیازمند به‌دست‌آوردن شباهت معنایی میان اسناد است. اندازه‌گیری دقیق شباهت معنایی بین واژه‌های موجود در اسناد، یک مشکل مهم در وب‌کاوی، بازیابی اطلاعات و پردازش زبان طبیعی است. برنامه‌های کاربردی وب‌کاوی نیاز به توانایی دقیق اندازه‌گیری شباهت معنایی بین مفاهیم یا موجودیت‌ها دارند [4]. روش‌های شباهت معنایی با استفاده از تعبیه لغت به‌عنوان یکی از عناصر تشکیل‌دهندهٔ بردار شباهت، هر متن را به‌عنوان یک بردار میانگین از لغات، نمایش می‌دهند [5]. *کمار*[4] *و همکاران* [6] یک رویکرد جدید را بر اساس شباهت کسینوسی[5] ارائه کرده‌اند. در این رویکرد ساختار معنایی مبتنی بر شباهت کسینوسی نه‌تنها وزن تقریبی بین واژگان را مشخص، بلکه همچنین روابط معناشناختی بین یک گروه از لغات را نیز محاسبه می‌کند. در نمایش بردار مبتنی بر یک سند، هر عنصر به تعداد هنجارسازی[6]شده‌ای از وقوع یک اصطلاح پایه در سند اشاره دارد. برای شمارش تعداد وقوع یک اصطلاح پایه، مدل کیف واژگان[7]، تطبیق دقیق واژگان را انجام می‌دهد که می‌تواند به‌عنوان نگاشت سخت از واژگان، به اصطلاح پایه محسوب شود. همچنین این روش‌ها توانایی ضبط معانی در پشت داده‌های متداول را ندارند. برای حل این مشکل *ژائو*[8] *و همکاران* [7] یک نگاشت فازی بر اساس همبستگی معنایی میان واژگان را پیشنهاد کرده‌اند. در این روش اندازه‌گیری شباهت بین ضمایم واژه توسط شباهت کسینوسی انجام شده است که در آن تطبیق معناشناختی به جای تطبیق دقیق رشتهٔ لغت استفاده می‌شود. ورودی پایه برای معانی وب‌معنایی ورودی کاربر است، اما دقت خروجی

[1] Filter
[2] Index
[3] Items
[4] Kumar
[5] Cosine Similarity
[6] Normalized
[7] Bag of Words
[8] Zhao

[DOI: 10.29252/jsdp.17.1.29]

داده‌ها بر اساس طبقه‌بندی دامنه است [6]. خوشه‌بندی در آمار و یادگیری ماشین، یکی از شاخه‌های یادگیری بدون نظارت است و فرآیندی است که در طی آن، نمونه‌ها به گروه‌های مشابه یکدیگر تقسیم می‌شوند که به این گروه‌ها خوشه گفته می‌شود؛ بنابراین هر خوشه مجموعه‌ای از اشیا است که در آن اشیا با یکدیگر مشابه بوده و با اشیای موجود در خوشه‌های دیگر غیرمشابه است. خوشه‌بندی اسناد یکی از مهم‌ترین روش‌های کاوش اسناد است. سامانه‌های رایج بازیابی اطلاعات و خوشه‌بندی اسناد بر واژگان کلیدی استوار هستند. با توجه به این‌که واژگان کلیدی مختلف می‌توانند برای توصیف یک مفهوم استفاده شوند، این سامانه‌ها می‌توانند نتایج نادرست و ناقصی را ایجاد کنند. همچنین شناسایی روابط معنایی موجود بین واژگان، نیاز به استخراج دانش دامنه موردنظر دارد.

در این مقاله یک الگوریتم خوشه‌بندی جدید برای افزایش دقت در جستجوی معنایی پیشنهاد می‌شود. در گام نخست داده‌های مرتبط با درخواست ارسالی کاربر با بهره‌گیری از الگوریتم (TF-IDF) وزن‌دهی می‌شوند. محاسبهٔ شباهت ابتدایی میان اسناد با استفاده از شباهت کسینوسی انجام می‌پذیرد. برای کشف روابط معنایی پنهان و همچنین بالابردن دقت جستجو مؤلفه‌هایی[1] نظیر نقش دستوری[2] و لغات مترادف به الگوریتم جستجو اضافه می‌شود. علاوه‌بر مؤلفه‌های نامبرده، مؤلفه‌های معنایی که به‌عنوان مهم‌ترین بخش در تشخیص روابط معنایی در نظر گرفته می‌شوند، استخراج خواهند شد. این مؤلفه‌ها به‌عنوان فراداده[3] در مجموعه داده‌های ما قرارگرفته‌اند. در گام بعدی مجموع این مؤلفه‌ها برای وزن‌دهی به‌عنوان ورودی یک زیرسامانه فازی ارسال می‌شوند؛ درنهایت با استفاده از سامانه تصمیم‌گیری فازی و استفاده از خوشه‌بندی سلسله‌مراتبی الگوریتمی با نام HFCS[4] معرفی می‌کنیم تا درصدد حل مقداری از مشکلات نام‌برده برآییم.

در سال‌های اخیر تلاش‌های بسیاری برای خوشه‌بندی داده‌های حجیم و کشف روابط معنایی میان آن‌ها صورت گرفته است. بر اساس بررسی‌های انجام‌شده درعمل روش مشابه دیگری وجود ندارد که با استفاده از خوشه‌بندی تجمعی و تصمیم‌گیری فازی امکان ارائهٔ بهترین پاسخ به جستجوی کاربر را بر روی داده‌های وب‌معنایی فراهم کند؛ علاوه‌بر موارد بالا، روش پیشنهادی برای کاربر، به‌صورت سلسله‌مراتبی استفاده از مشابه‌ترین اسناد وب را از لحاظ معنایی فراهم می‌کند. با استفاده از این دو روش، دانش عمیق‌تری از معانی لغات استفاده‌شده در اسناد وب مشتق شده و شباهت میان اسناد قابل اعتمادتر است. در روش HFCS به جای گروه‌بندی اسناد به فهرست‌های گسترده‌ای از نتایج، اسناد به خوشه‌های قابل‌فهم برای ماشین تبدیل می‌شوند که این امر در راستای غایت نهایی وب معنایی است.

ساختار ادامه مقاله بدین قرار است: در بخش بعدی به مروری بر سوابق و فعالیت‌های انجام‌شده درباره این موضوع پرداخته می‌شود؛ سپس جزئیات و نحوهٔ پیاده‌سازی روش پیشنهادی بیان می‌شود. در بخش بعد نتایج و تحلیل خروجی آن‌ها مورد بررسی می‌گیرد و نحوهٔ مقایسهٔ روش پیشنهادی با دیگر روش‌های موجود بیان و درنهایت از دستاوردهای روش پیشنهادی ما نام برده می‌شود.

۲- کارهای مشابه

در سال‌های اخیر برخی روش‌های خوشه‌بندی برای یافتن روابط معنایی پنهان میان واژگان معرفی‌شده‌اند که در ادامه به شرح برخی از آن‌ها پرداخته شده است.

در مقاله [8] مدلی برای طبقه‌بندی انواع الگوریتم‌های خوشه‌بندی ارائه شده است. خوشه‌بندی سلسله‌مراتبی یکی از روش‌های خوشه‌بندی است که به خوشه‌های نهایی بر اساس میزان شباهت آن‌ها ساختاری سلسله‌مراتبی تعلق می‌گیرد. این ساختار به‌صورت درختی ارائه می‌شود که هر سطح آن مراحل خوشه‌بندی را نمایش می‌دهد. به این درخت سلسله‌مراتبی دندوگرام[5] می‌گویند. روش‌های خوشه‌بندی بر اساس ساختار سلسله‌مراتبی تولیدی توسط آن‌ها به‌طورمعمول به دو دستهٔ زیر تقسیم می‌شوند:

۱. روش تجمعی[6]: این روش به روش پایین به بالا نیز معروف است. در روش تجمعی، ابتدا هر سند به‌عنوان یک خوشه در نظر گرفته می‌شود. در ادامه با به‌کارگیری یک الگوریتم تکرارشونده، در هر مرحله خوشه‌های دارای ویژگی‌های مشابه با یکدیگر ادغام‌شده و خوشهٔ جدیدی را تشکیل می‌دهند. این روند تا شکل‌گیری یک ساختار سلسله‌مراتبی از خوشه‌ها ادامه می‌یابد.

[1] Parameters
[2] Part of Speach
[3] MetaData
[4] Hierarchical Fuzzy Clustering Semantics

[5] Dendogram
[6] Agglomerative

[DOI: 10.29252/jsdp.17.1.29]

۲. روش تفکیکی[1]: از این روش به‌عنوان روش بالا به پایین نیز یاد می‌شود. در این روش ابتدا تمام اسناد به‌عنوان یک خوشه در نظر گرفته می‌شود؛ سپس با به‌کارگیری یک الگوریتم تکرارشونده، در هر مرحله داده‌ای که کمترین شباهت را با داده‌های دیگر دارد، به‌عنوان خوشه‌ای مجزا در نظر گرفته می‌شود. این کار ادامه می‌یابد تا یک ساختار سلسله‌مراتبی از خوشه‌ها ایجاد شود.

در مقاله [9] یک روش یک‌پارچه‌سازی فهرست‌های معنایی نهفته[2] به همراه خوشه‌بندی سلسله‌مراتبی با رویکرد تجمعی، پیشنهاد شده است. مشاهده می‌شود که پاسخ‌های جستجو با روش‌های معمول خوشه‌بندی نتایج[3]، راضی‌کننده نیستند و لغات می‌توانند به‌صورت بهتری خوشه‌بندی شوند. هنگامی‌که روش‌های اولیه خوشه‌بندی نتایج از مدل مجموعه‌ای از واژگان استفاده می‌کنند، پیداکردن روابط معنایی به‌صورت ساده میسر نخواهد بود. هدف از ترکیب این دو روش، پیداکردن مترادف‌ها و چندمعنایی‌هاست؛ همچنین این روش از نظر کیفیت خوشه‌ای بسیار قابل‌توجه است.

در مقاله [10] ایده‌ای برای استفاده از خوشه‌بندی مبتنی بر ویژگی در خوشه‌بندی اسناد ارائه شده است. با فرض وجود حجم انبوهی از داده‌ها، استفاده از تقسیم‌بندی موضوعی خوشه‌بندی را ساده‌تر خواهد کرد. خوشه‌بندی مبتنی بر ویژگی از الگوریتم میانگین[4] K برای خوشه‌بندی داده‌های متوالی استفاده کرده و ویژگی‌های اسناد را به‌عنوان دنباله‌ای از لغات ارائه می‌دهد. گاهی برای پردازش داده‌های متوالی، ویژگی‌ها باید از مجموعه‌ای از اسناد متنی بدون ساختار استخراج شوند؛ بنابراین، فرآیند پیش‌پردازش برای ارائهٔ شکل مناسب از ویژگی‌های سند ضروری است. برای این منظور دو نوع الگوی ترتیبی وجود دارد: توالی تکرار یک لغت[5] و بیشینه توالی تکرار[6] که هر دو نوع برای داده‌های متنی مناسب هستند. تفاوت در این دو الگو استفاده از بیشینه تکرارها در MFS به جای تمام تکرارها در FWS است؛ بنابراین تعداد مؤلفه‌های روش بیشینه در یک سند کمتر از تعداد مؤلفه‌های تکرار یک لغت است. در این مقاله از روش بیشینه تکرارها استفاده شده و نتایج نشان می‌دهد که دقت نتایج خوشه‌بندی در مجموعهٔ داده‌ها توسط تعداد مؤلفه‌ها در خوشه‌های مورد هدف مشخص می‌شود.

در مقاله [11] یک الگوریتم خوشه‌بندی سلسله‌مراتبی فازی با رویکرد تفکیکی ارائه شده است. این الگوریتم سلسله‌مراتبی تولید می‌کند که در آن، یک سند می‌تواند به چندین خوشه یا فرزندان یک خوشه تعلق داشته باشد. این خوشه‌ها در یک سطح سلسله‌مراتبی خاص‌تر از سطح بالا یا پایین قرار می‌گیرند. روش پیشنهادی برای تقسیم هر خوشه به خوشه‌های فرزند از یک الگوریتم تقسیم‌کننده، بر اساس روش K-Means به‌صورت اصلاح‌شده، بهره می‌برد. در هر مرحله برای تعیین تعداد مناسب خوشه‌ها از یک الگوریتم احتمالی فازی C-Means استفاده شده است. این الگوریتم تعداد خوشه‌های مناسب برای تولید در سطح نخست را بر اساس اندازهٔ یک بی‌نظمی[7] تعیین می‌کند و تصمیم می‌گیرد که آیا یک گره می‌تواند بر اساس تراکم[8] تقسیم شود.

در مقاله [12] تعدادی از روش‌های خوشه‌بندی مختلف مورد تجزیه و تحلیل قرارگرفته است؛ به‌خصوص خوشه‌بندی تجمعی و خوشه‌بندی تفکیکی. در این مقاله برای تجزیه و تحلیل اندازه‌گیری کیفیت خوشه‌ها از سه مؤلفهٔ: اندازه‌گیری سازگاری، گذشت زمان و شاخص سیلوئت[9] استفاده شده است. روش سیلوئت یکی از متداول‌ترین و بهترین روش‌های اعتبارسنجی الگوریتم‌های خوشه‌بندی است. این روش کیفیت خوشه را به‌صورت گرافیکی بررسی می‌کند. در این مقاله روش سیلوئت انتخاب شده است؛ زیرا این روش امکان نمایش خوشه‌های یک مجموعه‌داده با حالت پراکنده را دارد [13]. این شاخص مقدار تشابه یک شیء به اشیای درون خوشه را در مقایسه با شباهت آن شیء به اشیای خارج از خوشه اندازه‌گیری می‌کند [14]. نتایج اندازه‌گیری نشان می‌دهد که الگوریتم خوشه‌بندی تجمعی برای خوشه‌بندی اشیای داده مناسب‌تر است.

در مقاله [15] با توجه به گروه‌بندی سلسله‌مراتبی داده‌ها، به‌عنوان واژگان و جملات در اسناد متنی، یک روش یادگیری ساختاری برای استخراج موضوعات و مضامین[10] پنهان، میان جملات و واژگان ارائه شده است. در این روش رابطه بین مضامین و موضوعات، تحت گروه‌بندی‌های مختلف، از طریق یک روش بدون نظارت، بدون محدودکردن تعداد خوشه‌ها مورد بررسی قرار گرفته است. در این روش از فرآیند شکستن چوب[11] برای ارائه شباهت میان جملات

[1] Divisive
[2] Latent Semantic Indexing
[3] Search Result Clustering
[4] K-Means
[5] Frequent Word Sequence
[6] Maximal Frequent Sequence
[7] Entropy
[8] Density
[9] Silhouette
[10] Themes
[11] Stick-Breaking

[DOI: 10.29252/jsdp.17.1.29]

مختلف استفاده شده است. در این مقاله یک مدل موضوعی سلسله‌مراتبی ایجاد می‌شود که با استفاده از الگوریتم‌های فاقد مؤلفهٔ بیزین[1]، اسناد ناهمگن را نشان می‌دهد. این مدل در ساختن ساختار درخت معنایی برای جملات و واژگان مرتبط به کار می‌رود. همچنین برتری استفاده از مدل سلسله‌مراتبی درخت معنایی برای انتخاب جملات نمایش داده می‌شود.

در مقاله [16] به پیچیدگی‌های زمانی و فضایی موجود در روش استاندارد خوشه‌بندی سلسله‌مراتبی داده‌ها پرداخته شده است. به‌طور عمده این روش‌ها نیازمند محاسبه و ذخیره یک ماتریس فاصلهٔ زوج است. در این مقاله یک روش محاسبهٔ موازی در خوشه‌بندی سلسله‌مراتبی تجمعی برای به کمینه‌رساندن تعداد زوج فاصله‌های محاسبه‌شده بدون کاهش عملکرد خوشه‌ای ارائه شده است.

در مقاله [17] یک رویکرد جدید خوشه‌بندی سلسله‌مراتبی برای خودسازمان‌دهی اسناد علمی، ارائه‌شده است. روش‌های فعلی مبتنی بر واژگان کلیدی برای مدیریت محتوای سند زمانی که معانی جزئی از محتوای فنی یا تخصصی باشند، اغلب ناسازگار و بی‌اثر هستند؛ بنابراین، یک روش جدید خودکار برای تفسیر و خوشه‌بندی مدارک علمی با استفاده از یک هستی‌شناسی ارائه شده است. علاوه‌بر این، یک روش کنترلی با استفاده از منطق فازی برای مطابقت خوشه‌های مستقل و هستی‌شناسی مشتق‌شده از آن‌ها مورداستفاده قرارگرفته است. نتایج نشان می‌دهد که رویکرد خوشه‌بندی فازی مبتنی بر هستی‌شناسی از رویکرد K-Means در دقت، جامعیت، آزمون F و بی‌نظمی شانون[2] برتر است.

در مقاله [18] یک چارچوب[3] نوین از متغیرهای مارکوف به نام توزیع احتمالی شرطی وزن‌دار[4] پیشنهاد و از این چارچوب برای مدل‌کردن، طبقه‌بندی خوشه‌ها استفاده شده است. برای مدل‌سازی ابتدایی، از مدل مارکوف مرتبه نخست بر اساس یک بردار از شاخص‌های فازی استفاده شده که این روش شاخص وزن‌دهی فازی[5] نامیده شده است؛ درنهایت بر اساس یک چارچوب بهینه‌سازی آبشاری که ترکیبی از مدل‌های توزیع احتمالی شرطی وزن‌دار و شاخص وزن‌دهی فازی است، یک الگوریتم خوشه‌بندی سلسله‌مراتبی جدید با وزن فازی ارائه می‌شود.

در مقاله [1] برای کشف معانی پنهان در اسناد وب، یک الگوریتم بر اساس فضای توپولوژیکی تشکیل‌شده از متغیرهای زبانی فازی ارائه شده است. الگوریتم پیشنهادی ویژگی‌ها را از اسناد وب با استفاده از روش فضای تصادفی شرطی[6] استخراج و یک فضای توپولوژیکی زبانی فازی را بر اساس ترکیب ویژگی‌ها ایجاد می‌کند. در این روش برای محاسبه شباهت از الگوریتم تجزیه و تحلیل معنایی نهفته[7] استفاده شده است. این روش با بهره‌گیری از خوشه‌بندی فازی وابستگی و هماهنگی ویژگی‌های هر مجموعه، سلسله‌مراتبی از مجتمع‌های معنایی متصل به نام مفاهیم را به‌وجود می‌آورد. در اینجا ارزیابی مفاهیم به دو روش: ارتباط یک سند متعلق به یک موضوع و تفاوت بین موضوعات دیگر انجام گرفته شده است.

در مقاله [19] روشی جدید برای طبقه‌بندی اسناد نیمه‌ساختاری با استفاده از سامانه مبتنی بر قاعده فازی ارائه شده است. ایده این رویکرد تقسیم صفحات وب به بخش‌های معنایی مختلف است. در این روش با استفاده از ویژگی‌های سامانه منطق فازی و اختصاص وزن به اصطلاحات، روشی جدید برای نمایش اسناد نیمه‌ساختاری ارائه می‌شود.

در مقاله [1] از روش فضای تصادفی شرطی برای استخراج ویژگی‌ها استفاده شده است. روش فضای تصادفی شرطی یک الگوریتم یادگیر و هزینهٔ محاسباتی در مرحله آموزش این الگوریتم زیاد است. همچنین این امر باعث می‌شود که با به‌روزرسانی یا تغییر شکل اطلاعات، فرآیند یادگیری این الگوریتم بسیار دشوار باشد؛ به همین دلیل در روش پیشنهادی از چندین هستان‌شناسی پویا برای استخراج ویژگی‌ها استفاده کرده‌ایم. علاوه‌براین در این روش عامل شباهت، شباهت هر سند با سند دیگر در نظر گرفته شده است؛ برای تحقق امر باید هر سند با تمام اسناد موجود در مجموعه داده مقایسه شود و به‌ازای هر مقایسه باید میان آن دو سند هنجارسازی انجام گیرد؛ به همین دلیل در روش HFCS عامل شباهت را به شباهت میان عبارت مورد نظر کاربر با اسناد موجود در مجموعه داده تغییر دادیم در این حالت تنها نیازمند مقایسهٔ سند با عبارت مورد نظر کاربر خواهیم بود. در این حالت هنجارسازی را به‌ازای هر سند به دو مرحله کاهش می‌دهیم.

[1] Bayesian
[2] Shannon
[3] Framework
[4] Weighted Conditional Probability Distribution
[5] Weighted Fuzzy Indicator
[6] Conditional Random Field
[7] Latent Semantic Analysis

[DOI: 10.29252/jsdp.17.1.29]

۳- منطق فازی

منطق فازی یک نوع منطق است که روش‌های متنوع نتیجه‌گیری در مغز بشر را جایگزین الگوهای ساده‌تر ماشینی می‌کند. درواقع می‌توان این‌طور استدلال کرد که مغز بشر به ورودی‌های اطلاعاتی دقیق نیازی ندارد، بلکه قادر است تا کنترل تطبیقی را به‌صورت بالایی انجام دهد و این در مورد ماشین نیز صادق است. ساده‌ترین تلقی برای تعریف منطق فازی این است که منطق فازی جواب یک سؤال را به‌جای تقسیم به دو بخش درست یا نادرست، در اصل به یک محدوده جواب، دراین‌بین توسعه داده است. نمونه معمول آن، وجود رنگ خاکستری در طیف رنگی بین سیاه و سفید است؛ اما دایره عمل منطق فازی، از این هم گسترده‌تر است و می‌توان با استفاده از قواعد منطق فازی، جواب‌های فازی متناسب با پرسش را ارائه کرد.

منطق فازی می‌کوشد، دانستنی‌هایی که تا حدودی قابل توصیف و بیان زبان‌شناختی بوده، ولی امکان کمّی کردن آن‌ها با کمک ریاضیات سنتی به‌طورمعمول وجود ندارد را به‌صورتی منظم، منطقی و ریاضیاتی از طریق متغیرهای زبانی با یکدیگر هماهنگ گرداند. متغیرهای زبانی، متغیرهایی هستند که مقادیرشان اعداد نیستند، بلکه لغات یا جملات یک زبان طبیعی یا ساختگی هستند. یک متغیر زبانی درواقع، یک عبارت زبان طبیعی است که به یک مقدار کمیت خاص اشاره دارد و دراصطلاح مانند مترجم عمل می‌کند و به‌کمک تابع عضویت، نشان داده می‌شود؛ تابع عضویت به شما اجازه می‌دهد تا یک واژه زبانی را کمی‌سازی کرده و یک مجموعه فازی را به‌صورت نموداری نمایش دهید.

اگرچه سامانه‌های فازی پدیده‌های غیرقطعی و نامشخص را توصیف می‌کند بااین‌حال نظریه فازی یک نظریه دقیق است.

۴- ارائه روش پیشنهادی HFCS

همان‌طور که در بخش‌های قبل بیان شد خوشه‌بندی فرایندی است که در نتیجهٔ آن گروهی از خوشه‌ها بر اساس معیار شباهت ایجاد می‌شوند؛ بنابراین معیارهای شباهت نقش مهمی را در خوشه‌بندی ایفا می‌کنند. یکی از روش‌های رایج برای ارزیابی شباهت بین الگوها استفاده از معیار فاصله است. بیشترین معیار فاصله مورد استفاده در الگوریتم‌های مختلف استفاده از فاصله اقلیدسی است. فاصله اقلیدسی به‌طور معمول برای خوشه‌بندی مجموعه‌داده با ابعاد زیاد کارآمد نیست؛ زیرا ویژگی‌ها با مقیاس بالا ویژگی‌های کوچک را مغلوب می‌کنند. این مشکل می‌تواند با هنجارکردن ویژگی‌ها برطرف شود. یکی از روش‌ها برای انجام این کار استفاده از شباهت کسینوسی است.

در این مقاله با توجه به چالش‌های موجود در جستجو و رتبه‌دهی میان اسناد به‌منظور بالابردن دقت جستجو در اسناد وب‌معنایی یک روش خوشه‌بندی جدید با نام HFCS[1] ارائه شده است. روش‌های معمول جستجو برای جستجو بر اساس واژگان کلیدی در گذشته مناسب به نظر می‌رسید، اما با حجم روزافزون اطلاعات، روش‌های معمول به‌تنهایی کارآمد نخواهد بود؛ همچنین با فرض کارآمدی این روش‌ها، امکان طبقه‌بندی و نمایش بهترین نتیجه برای جستجو به کاربر وجود نخواهد داشت. همان‌طور که در بخش‌های قبل توضیح داده شد، روش ارائه شده برای حل این مشکل، استفاده از وب‌معنایی است. اساس جستجوی معنایی، روابط میان واژگان است؛ اما در جستجوی معنایی برای کشف این روابط مستلزم استخراج ویژگی‌های موردنیاز از اسناد هستیم. در عملیات پردازش متن، انتخاب یک روش برای استخراج ویژگی‌ها یا اصطلاحات، یک مشکل چالش‌برانگیز است. رویکردهای مختلفی برای انتخاب ویژگی و اختصاص وزن در روش‌های پردازش متن وجود دارد. رویکرد محبوب برای استخراج ویژگی‌ها مدل فضایی برداری[2] است؛ که هدف از آن ارائه اسناد به‌عنوان یک بردار است. افزونه‌های زیادی برای مدل فضایی برداری تولید شده‌اند که از مهم‌ترین آن‌ها می‌توان تجزیه و تحلیل معنایی نهفته یا تجزیه و تحلیل معنایی نهفته‌ی احتمالی[3] را نام برد که برخلاف سایر روش‌های فضای برداری مجموعه‌ای از مفاهیم مرتبط با اسناد و اصطلاحات آن را تولید می‌کنند. ایده اصلی این روش‌ها رخ‌دادن روابط پنهان، میان اصطلاحات نزدیک‌تر در اسناد مشابه، ممکن است. پژوهش‌گران روش‌های استخراج اطلاعات را به دو روش: تحت نظارت[4] و بدون نظارت[5] تقسیم‌بندی کرده‌اند. هر دو روش استخراج ویژگی تحت نظارت و بدون نظارت از هر اصطلاح به‌عنوان ابعادی از یک مدل فضایی برای نشان‌دادن یک سند در یک واحد خاص استفاده می‌کنند. یکی از شایع‌ترین راه‌های بدون نظارت در

[1] Hierarchical Fuzzy Clustering Semantics
[2] Vector Space Model
[3] Probabilistic Latent Semantic Analysis
[4] Supervised
[5] Unsupervised

[DOI: 10.29252/jsdp.17.1.29]

وزن‌دهی مدارک روش (TF-IDF) است؛ که در این مقاله از این روش به‌منظور وزن‌دهی اولیه اسناد استفاده شده است. تلاش‌های بسیاری برای اندازه‌گیری شباهت بین اسناد مبتنی بر مدل فضایی برداری مانند شباهت کسینوسی، شباهت اقلیدسی[1]، ضریب جاکارد[2] و دیگر روش‌ها انجام‌شده است. در روش پیشنهادی ما، برای محاسبه شباهت اولیه میان اسناد از ایده بردار اصطلاحات به‌همراه روش شباهت کسینوسی استفاده شده است. ساختار تولیدشده بر اساس معماری وب‌معنایی در لایهٔ توصیف منابع[3] از راه فرادادهٔ افزوده‌شده به متن صفحات در طبقه‌بندی اسناد کمک شایانی می‌کند. با استفاده از فراداده بخش‌های مختلف اسناد بر اساس اهمیت آن‌ها طبقه‌بندی می‌شوند. پس از طبقه‌بندی یک سند به بخش‌های مختلف، با یک سامانه تصمیم‌گیری انسانی می‌توان به هر بخش از سند یک وزن خاص اختصاص داد. سامانه مبتنی بر قاعده فازی می‌تواند از دانش تخصصی فرد برای ساخت یک سامانه مبتنی بر قاعده استفاده کند. این سامانه از لحاظ معنایی شباهت و ارتباط میان اسناد را مشخص خواهد کرد. در این سامانه برای افزایش دقت در کشف روابط معنایی مؤلفه‌های مختلفی به سامانه اضافه خواهد شد. به‌ازای هر مؤلفهٔ افزوده‌شده به سامانه، محاسبهٔ شباهت میان اسناد انجام گرفته و مجموع این شباهت‌ها برخلاف دیگر روش‌های موجود به‌عنوان ورودی سامانه فازی قرار خواهند گرفت. پس از وزن‌دهی سامانه فازی، خروجی سامانهٔ جستجو با استفاده از وزن شباهت اولیه و وزن خروجی از سامانهٔ فازی با استفاده از خوشه‌بندی سلسله‌مراتبی رتبه‌بندی خواهد شد و بر اساس این رتبه خروجی جستجو به‌ترتیب نمایش داده‌شده و اسناد مرتبط از لحاظ معنایی به کاربر پیشنهاد می‌شود.

در بخش بعدی به توضیح دقیق‌تر و موشکافانه راجع به مراحل نامبرده خواهیم پرداخت.

۱-۴- پیاده‌سازی روش HFCS

همان‌طورکه مشاهده می‌شود در شکل (۱) مراحل کلی جستجو و رتبه‌دهی اسناد تا بازگشت پاسخ به کاربر نمایش داده شده است.

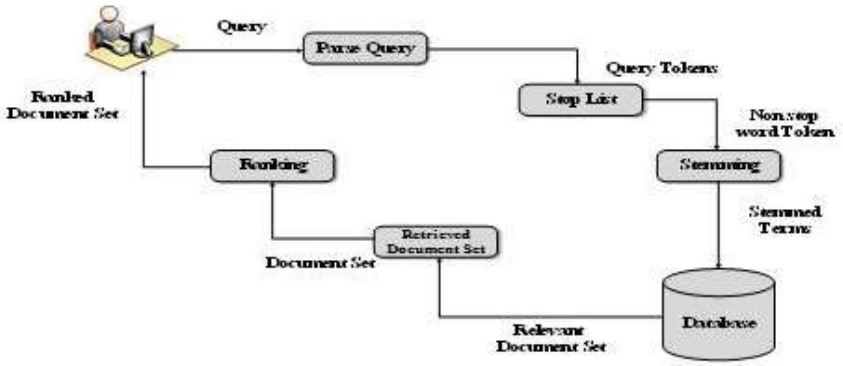
در گام نخست جستجوی معمول هنگامی‌که کاربر جستجوی موردنظر خود را ارسال می‌کند، عبارت جستجوشده از طرف کاربر تحت فرآیند تحلیل واژگانی[4] به لغات جدا تبدیل می‌شود برای مثال: زمانی که کاربر عبارت: «کتاب صدسال تنهایی از مارکز» را جستجو می‌کند، فهرستی از لغات به‌صورت [کتاب، صد، سال، تنهایی، از، مارکز] تشکیل‌شده و به‌ازای هریک از این لغات در صورت تکراری‌نبودن تمام اسناد موجود جستجو خواهد شد. پس از انجام جستجو متوجه می‌شویم که نتیجهٔ جستجو مطلوب نخواهد بود؛ زیرا در همین مثال لغت "از" در بسیاری از متون فارسی موجود است و زمانی که جستجو انجام می‌شود تمام اسناد شامل آن بازیابی خواهد شد. به این‌گونه واژگان ایست‌واژه[5] گفته می‌شود. ایست‌واژه‌ها لغاتی هستند که برخلاف تکرار فراوان در متن، ازلحاظ معنایی دارای اهمیت پایینی هستند، مانند "اگر"، "و"، "ولی"، "که" و غیره. در نگاه اولیه واژگان ربط و تعریف، ایست‌واژه به نظر می‌آیند؛ اما بسیاری از افعال، افعال کمکی، اسم‌ها، قیدها و صفات نیز ایست‌واژه شناخته‌شده‌اند. در اغلب کاربردهای متن، حذف این واژگان، نتایج پردازش را به‌شدت بهبود می‌دهد و سبب کاهش بار محاسبات و افزایش سرعت خواهد شد. به همین دلیل این واژگان اغلب در مرحلهٔ پیش‌پردازش، حذف می‌شوند. برای زبان فارسی چندین فهرست از این واژگان منتشر شده است که به‌طور میانگین شامل پانصد واژه است. پس از حذف ایست‌واژه‌ها، در مرحلهٔ بعد اقدام به ریشه‌یابی[6] لغات می‌کنیم.

ریشه‌یابی عبارت است از به‌دست‌آوردن ریشه واژگان با حذف پسوندها و پیشوندها به‌طوری‌که واژگان با ریشه یکسان دارای شکل یکسان شوند. معمول‌ترین هدف ریشه‌یابی استفادهٔ از پایه و اساس یک کلمه، برای شناخت واژگان مشابه است. در سامانه بازیابی اطلاعات استفاده از واژگان ریشه‌یابی‌شده به‌جای واژگان اصلی می‌تواند کارایی سامانه را ارتقا دهد. بعد از انجام این مراحل، سامانه فهرستی از اسناد موردنظر کاربر را تشکیل خواهد داد؛ اما برای نمایش بهترین نتیجه به کاربر و طبقه‌بندی اسناد نیازمند سازوکاری برای رتبه‌دهی به این اسناد هستیم. در این روش رتبه‌دهی شامل محاسبهٔ شباهت، اختصاص مؤلفه‌های افزایش دقت و اختصاص وزن در سامانه فازی است که در ادامه به تشریح هریک خواهیم پرداخت.

[1] Euclidean Similarity
[2] Jacquard Coefficient
[3] Resource Description Framework (RDF)

[4] Tokenization
[5] Stop word
[6] Stemming

[DOI: 10.29252/jsdp.17.1.29]



(شکل-۱): فلوچارت جستجو در اسناد
(Figure-1): Search System FlowChart

۲-۴- محاسبه شباهت

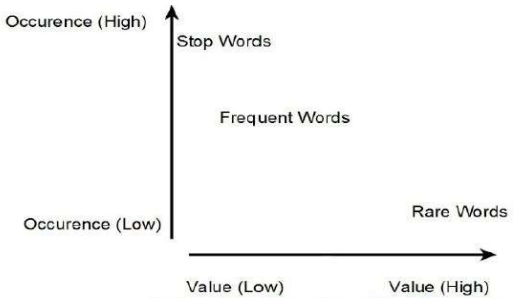
همان‌طور که گفته شد، با افزایش تعداد واژگان در جستجو نیازمند سازوکاری برای طبقه‌بندی اسناد و نمایش بهترین نتیجه به کاربر هستیم، مدل فضای برداری یک چارچوب ساختاریافتهٔ ریاضی با مجموعه‌ای از عناصر[1] است که مجموعهٔ این عناصر بردار نامیده می‌شوند و هر بردار می‌تواند به‌وسیلهٔ مقادیر، نمایش و گسترش داده شود. درواقع با استفاده از مدل فضای برداری می‌توان یک متن یا محتوا را به حالت یک بردار (ماتریس) تبدیل کرد و عملیات ریاضی روی آن انجام داد. باید توجه داشت یک بردار می‌تواند به‌صورت مجموعه‌ای از عناصر نمایش داده شود.

$$V = a_1 v_1 + a_2 v_2 + \cdots + a_k v_n \qquad (۱)$$

در اینجا v_n نشان‌دهندهٔ عناصر و a_k برابر وزن آن‌هاست. برای محاسبهٔ وزن هریک از عناصر از الگوریتم TF-IDF استفاده شده است. در این شیوه، به لغات یک وزن بر اساس فراوانی آن در سند اختصاص داده می‌شود. درواقع این سامانه وزن‌دهی، اهمیت یک واژه در یک سند را نشان می‌دهد. این مسأله کاربردهای بسیاری در بازیابی اطلاعات دارد. وزن لغت با افزایش تعداد تکرار آن در متن افزایش می‌یابد، اما توسط تعداد لغات در متن کنترل می‌شود، چراکه می‌دانیم در صورت زیادبودن طول متن، بعضی از لغات به‌طول طبیعی بیشتر از دیگران تکرار خواهند شد، اگرچه چندان اهمیتی در معنی نداشته باشند. ارزش TF-IDF از دو تابع TF[2] و IDF[3] به دست می‌آید که در آن TF نشان‌دهنده بسامد یک واژه در یک سند (تعداد تکرار یک واژه در یک سند) و IDF نشان‌دهنده بسامد معکوس یک واژه در سند است (تعداد تکرار یک واژه در کل اسناد). وزن TF-IDF از طریق رابطهٔ زیر محاسبه می‌شود.

$$W_{TF\text{-}IDF} = (1+\text{Log TF}) \times \text{Log}\left(\frac{N}{df_i}\right) \qquad (۲)$$

در اینجا N نشان‌دهندهٔ کل اسناد موجود و DF_i نشان‌دهندهٔ فهرست اسناد شامل واژهٔ مورد نظر است.



(شکل-۲): نمودار اهمیت لغات در روش TF-IDF [20]
(Figure-2): Graph of the Importance of Words in the TF-IDF Method [20]

روش‌های متعددی برای اندازه‌گیری شباهت میان دو سند ارائه شده است؛ اما با توجه به تبدیل هر سند به بردار در مرحله قبل توسط الگوریتم TF-IDF اندازه‌گیری شباهت میان بردارها بر اساس زاویهٔ ایجادشده میان آن‌ها، نسبت به دیگر روش‌ها بسیار ساده‌تر و کارآمدتر است. شباهت کسینوسی با ضرب داخلی بردارها تقسیم بر ضرب خارجی آن‌ها محاسبه می‌شود. این عبارت برابر است با کسینوس زاویهٔ میان آن‌هاست که در رابطهٔ (۳) نمایش داده شده است.

$$\text{Similirity (A, B)} = \frac{A.B}{||A||\,||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \qquad (۳)$$

[1] Elements
[2] Term Frequency and weighting
[3] Inverse Document Frequency

[DOI: 10.29252/jsdp.17.1.29]

در رابطهٔ (۳): مقادیر A و B دو سند هستند که قرار است شباهت میان آن‌ها سنجیده شود و مقدار i نشان‌دهندهٔ ویژگی استخراج‌شده در سند است که در الگوریتم TF-IDF مقداردهی شده است. با این روش می‌توان شباهت اولیه میان اسناد را محاسبه کرد؛ اما برای افزایش دقت آن نیازمند مؤلفه‌های دیگری هستیم که در ادامه به توضیح آن‌ها خواهیم پرداخت.

۴-۳- وزن‌دهی بر اساس نقش دستوری

با پردازش هر جمله می‌توان نقش دستوری یک لغت در آن جمله را مشخص کرد و به‌ازای آن یک وزن به آن لغت اختصاص داد. عملیات مشخص‌کردن نقش بر اساس یک سری برچسب‌های استانداردشده انجام می‌پذیرد. استفاده از این نوع وزن‌دهی یا برچسب دستوری می‌تواند در افزایش دقت جستجو مفید باشد؛ به‌خصوص برای واژگانی که طرز نوشتن یکسان ولی معنای متفاوتی دارند مانند: لغت book که اگر در جمله به‌عنوان اسم مورداستفاده قرار گیرد به معنای کتاب و اگر به‌عنوان فعل مورد استفاده قرار گیرد به معنای رزروکردن است. در بیش‌تر سامانه‌های جستجو ازجمله موتور جستجوی گوگل این برچسب‌ها به دو بخش: برچسب‌های نام و دیگر نقش‌های دستوری تقسیم‌بندی می‌شوند و میزان بسامد آن‌ها در الگوریتم TF-IDF تأثیر داده می‌شود [21]؛ اما با این تفسیر ما استفادهٔ معنایی از این برچسب‌ها را از دست خواهیم داد و برچسب‌هایی که برچسبی غیر از نقش «اسم» دارند با ارزش یکسان تلقی خواهند شد. علاوه‌بر این اگر از روش بالا برای نقش دستوری استفاده شود امکان تأثیر مؤلفه‌های متعدد در وزن TF-IDF میسر نخواهد بود. در روش پیشنهادی، برچسب‌های معنایی تنها در دو گروه قرار نگرفته‌اند. در این روش به‌جای شمارش تمام برچسب‌ها از برابری نقش دستوری استفاده شده است. برای این منظور ابتدا نیازمند یک لغت‌نامهٔ معنایی شامل برچسب‌های دستوری استانداردشده و همچنین سامانه‌ای برای برچسب‌زنی اسناد خواهیم بود؛ برای این منظور از کتابخانهٔ «POSTagger» برای برچسب‌زنی استفاده شده است. بعد از به‌دست‌آوردن نقش دستوری نیازمند روشی برای وزن‌دهی نقش دستوری به هریک از اسناد هستیم. برای محاسبه این وزن یک زیرسامانه[1] فازی طراحی شده است که در بخش بعدی به توضیح دربارهٔ آن خواهیم پرداخت.

[1] Sub System

۴-۴- زیرسامانهٔ وزن‌دهی فازی

به‌ازای هر مؤلفه‌ای که برای بالابردن دقت به الگوریتم جستجو اضافه می‌شود نیازمند سامانه‌ای هستیم که توانایی اختصاص یک وزن خاص به آن مؤلفه را دارد. برای این منظور در روش پیشنهادی ما یک زیرسامانه با بهره‌گیری از منطق فازی تعبیه شده است. با این روش دیگر محدودیتی در افزایش تعداد مؤلفه‌ها نخواهیم داشت و تنها به‌ازای هر مؤلفه نیازمند اضافه‌کردن قوانین جدیدی هستیم. همچنین با داشتن چنین زیرسامانه‌ای دیگر برای استفاده از مؤلفه‌های مختلف نیازمند هنجارسازی نخواهیم بود؛ چون تمامی مقادیر توسط سامانهٔ فازی تولید می‌شوند. برای پیاده‌سازی این زیرسامانه فازی نیازمند کتابخانه‌ای هستیم که امکان استفاده از منطق فازی و ایجاد قوانین موردنظر را برای ما فراهم کند برای این منظور از کتابخانهٔ «FuzzyLogicLibrary» که معروف به «AI Fuzzy» نیز هست، استفاده شده است.

```
Name: DefinedFuzzySubSystemRules
Input: Fuzzy Input Terms [ ] IT,
 Fuzzy Output Terms [ ] OT,
 Fuzzy Linguistic Input threshold [ ] ITH,
 Fuzzy Linguistic Output threshold [ ] OTH,
 inputTitle, outputTitle
Output: FuzzyRules
1: IR ← Null                    ◁InputRules
2: OR ← Null                    ◁OutputRules
3: Rules← Null
3: IR ← FuzzyInputVariable (inputTitle, 0.0, 1.0);
4: for (i = 0; i < Count (IT); i++) do
5:          IR.Add (IT, FuzzyType (ITH 1, ITH 2, …, ITH n))
            ◁IR.Add ("inputTerm1",
TriangularMembershipFunction(0.39, 0.5, 0.61))
6: end for
7:  OR ← FuzzyOutputVariable (outputTitle, 0.0, 1.0);
8: for (j = 0; j < Count (OT); j++) do
9:          OR.Add (T, FuzzyType (OTH 1, OTH 2, …,
OTH m))
10: end for
11: Rules ← ParseRule (if (inputTitle is ITHn) then outputTitle
is OTHm)
◁All Rules are defined as the same
12: FuzzyIntelligence.add (Rules)
```

(شکل-۳): شبه کد تعریف قوانین در زیرسامانه فازی

(Figure-3): Pseudo Code of the Defined Fuzzy SubSystem Rules

در بخش قبل به هریک از واژگان موجود در هر سند یک نقش دستوری اختصاص داده شد. حال با مقایسهٔ نقش دستوری واژگان موجود در هر سند با واژگان موجود در عبارت جستجو می‌توان تشخیص داد که چه تعداد از این نقوش باهم برابرند. همان‌طور که در ابتدا گفته شد در این الگوریتم از روش TF-IDF استفاده شده است. با داشتن تعداد

[DOI: 10.29252/jsdp.17.1.29]

تکرار واژگان یا وزن TF می‌توان درصدد صحیح نقوش دستوری را به‌ازای واژگان موجود در هر سند محاسبه و ماتریس نقش دستوری به‌ازای هر سند شکل داد؛ سپس با قراردادن این ماتریس به‌عنوان ورودی زیرسامانه فازی می‌توان، وزن فازی مؤلفهٔ نقش دستوری یا هر مؤلفهٔ دیگر از این قبیل را محاسبه کرد.

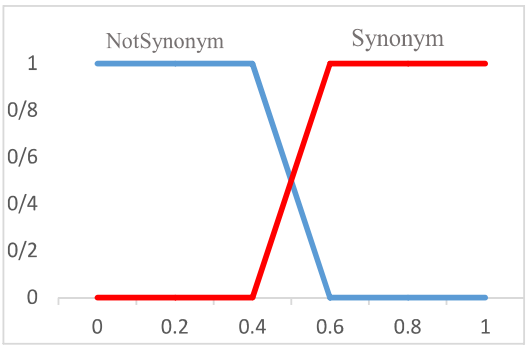
Name: FuzzySubSystem
Input: Primary Similarity Scores [ ] PSC, Primary Similarity Scores Variable [DocumnetNumber, Title, Body, Part of Speech, Synounym] PSCV, outputTitle
Output: Fuzzy Linguistic Variable

```
FSR [ ] ← Null                    ◁FuzzySubSystemRules;
SSInput [ ] ← Null                ◁SubSystemInput
SSOutput [ ] ← Null               ◁SubSystemOutput
FIV ← Null        ◁FuzzyInputVariable (Title, min, max)
FOV ← Null        ◁FuzzyOutputVariable (Title, min, max)
Result [ ] ← Null
if FSR = Null then
        FSR ← LoadFuzzySubSystemRules( );
        ◁ (see DefiendFuzzySubSystemRules(Table…))
else
        for (j = 1; j < Count (PSCV); j++) do
        ◁Count = 5
                FIV ← FSR.SelectInputByTitle (PSCV[ j ]);
                FOV ← FR.SelectOutputByTitle (outputTitle);
                SSInput.add (FIV, PSC[j ]);
                SSOutput ← CalculateFuzzyWeight (SSInput);
                Result [j-1] ← SSOutput [FOV]
        end for
end if
Return Result
```

(شکل-۴): شبه کد الگوریتم زیرسامانه فازی
(Figure-4): Pseudo Code of the Fuzzy SubSystem Algorithm

۴-۵- لغات مترادف

یکی دیگر از مؤلفه‌هایی که در افزایش دقت جستجو مؤثر است؛ لغات مترادف هستند. بر اساس اعلام W3C مؤلفه‌های نقش دستوری و لغات مترادف به‌عنوان بخشی از استاندارد SKOS[1] در ایجاد فراداده معرفی شده‌اند. این استاندارد به‌عنوان سامانه سادهٔ ساختاردهی دانش در وب‌معنایی معرفی شد. این استاندارد علاوه‌بر نقش دستوری و لغات مترادف شامل مؤلفه‌هایی از قبیل طبقه‌بندی[2]، رده‌بندی مردمی[3]، سرموضوع[4] (عنوان‌های مختلفی که یک موضوع را تشریح یا عنوان‌های یکسانی که موضوعات مختلفی را تشریح می‌کنند) و غیره است. در این مقاله با استفاده از کتابخانهٔ «NHunspell» مترادف‌ها به‌ازای لغات پراهمیت موجود در درخواست کاربر تشخیص داده می‌شوند. اهمیت این لغات بر اساس روش TF-IDF مشخص می‌شود. پس از استفاده از این کتابخانه، فهرستی سلسله‌مراتبی از لغات مترادف تشکیل خواهد شد که هرچه از ریشه دورتر می‌شویم اولویت لغات کاهش می‌یابد. درصورتی‌که جستجو به‌ازای تمام این لغات انجام گیرد، نتیجهٔ جستجو مطلوب نخواهد بود. برای حل این مشکل تنها لغات پراهمیت در جستجو در نظر گرفته می‌شوند؛ زیرا ممکن است هر لغت تعداد زیادی مترادف داشته باشد در این صورت اسناد بازیابی‌شدهٔ زیادی خواهیم داشت که بی‌ارزش هستند و به‌جای افزایش دقت در جستجو از دقت آن نیز می‌کاهند. پس از تشخیص اهمیت لغات این لغات، نیز در اسناد جستجو شده اما با درنظرگرفتن این‌که آن‌ها واژه‌های اصلی نیستند، درواقع به آن‌ها اولویت کمتری نسبت داده می‌شود. پس از بازیابی اطلاعات موردنظر از اسناد به‌ازای هر سند یک ماتریس تولید خواهد شد. در این بخش این ماتریس همانند مؤلفهٔ نقش دستوری به‌عنوان ورودی به زیرسامانه فازی تحویل داده شده و تابع عضویت فازی به‌ازای مؤلفهٔ واژگان مترادف تولید و وزن فازی به آن اختصاص داده خواهد شد.



(شکل-۵): تابع عضویت برای متغیر لغات مترادف
(Figure-5): Membership function for the linguistic variable Synonym tag

۴-۶- وزن‌دهی بر اساس برچسب‌های معنایی

در این بخش برای رتبه‌بندی میان اسناد بازیابی‌شده از برچسب‌های معنایی بهره‌گیری شده است. هر سند رویترز دارای برچسب‌های مشخصی است که بر اساس تعریف وب‌معنایی می‌توان این برچسب‌ها را به‌عنوان فراداده در نظر گرفت و بر اساس آن جستجو را انجام داد. در روش HFCS به جای محاسبهٔ شباهت کسینوسی به‌ازای هر سند، شباهت کسینوسی به‌ازای هر برچسب معنایی محاسبه خواهد شد.

[1] Simple Knowledge Organization System
[2] Classification
[3] Taxonomies
[4] Subject-Heading

[DOI: 10.29252/jsdp.17.1.29]

هریک از این برچسب‌ها می‌تواند به‌عنوان مؤلفه‌ای در افزایش دقت جستجو ایفای نقش کند. همچنین با این روش می‌توان میان برچسب‌ها اولویت ایجاد کرد، برای مثال زمانی که نتیجهٔ جستجو در عنوان یک سند یافت شد (در برچسب Title) اولویت بیشتری نسبت به یافت‌شدن نتیجه در متن خبر (در برچسب Body) داشته باشد یا می‌توان به واژگان کلیدی در انتهای هر سند اولویت بیشتری اختصاص داد. علاوه‌بر امکانات نام‌برده با این روش می‌توان به‌ازای رده‌بندی مردمی[1] یا دیگر رتبه‌بندی‌ها نیز یک برچسب اختصاص داد و مقدار به‌دست‌آمده را درنتیجهٔ جستجو اعمال کرد؛ درنهایت برای یکسان‌سازی داده‌ها و تبدیل آن به وزن فازی، میزان شباهت به‌دست‌آمده از هر برچسب معنایی، به‌عنوان مؤلفهٔ ورودی زیرسامانه فازی در نظر گرفته می‌شود و وزن فازی متناظر با آن محاسبه خواهد شد.

۷-۴- سامانه وزن‌دهی فازی

در روش پیشنهادی میزان شباهت به‌ازای هر برچسب معنایی محاسبه شد؛ اما بر اساس مشکلات نامبرده امکان استفاده از تمام این برچسب‌ها به‌صورت یکجا وجود نداشت. برای این منظور یک سامانه فازی طراحی شده است تا عمل رتبه‌دهی و تجمیع تمام این مؤلفه‌ها و درنهایت تصمیم‌گیری برای نمایش بهترین نتیجهٔ جستجو به کاربر را انجام دهد.

Name: Hierarchical Fuzzy Clustering Semantics (HFCS)
Input: Primary Similarity Scores [ ] PSC, Primary Similarity Scores Variable [DocumnetNumber, Title, Body, Part of Speech, Synounym] PSCV, Cosine Similarity CS, Number of Documents ND
Output: Hierarchical Cluster

```
1: FSSW [ ] ← Null
        ◁FuzzySubSystemWeight
2: FW ← Null
        ◁FuzzyWeight
3: Clusters[ ] ← Null
4: LoadFuzzyRules();
        ◁Laod Rules From Table-1
5: for (i = 0; i < ND; i++) do
6:      FSSW ← FuzzySubSystem (PSC, PSCV, outputTitle)   ◁PSC [1, 0.92, 0.84, 0.64, 0.001 ]
7:      FW ← CalculateFuzzyWeight (FSSW);
8:      Clusters.add (i, FW, CS);
9: end
10: NC ← Number of Cluster
11: while NC != 1 do
12:     MergeTwoClosetCluster (Clusters);
13:     NC ← NC - 1
14: end while
```

(شکل-۶): شبه‌کد الگوریتم HFCS
(Figure-6): Pseudo Code of the HFCS Algorithm

[1] Taxonomy

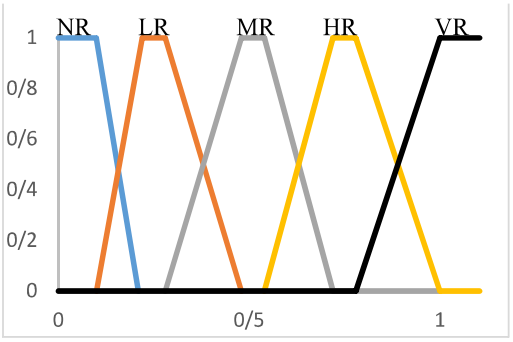
در بخش‌های قبل مؤلفه‌های افزایش دقت به‌عنوان ورودی زیرسامانه فازی در نظر گرفته شد و به‌ازای آن وزن فازی متناظر با آن محاسبه شد. در این مرحله برای استفاده از این سامانه فازی تمام مؤلفه‌های محاسبه‌شده به‌عنوان ورودی سامانه فازی منظور می‌شوند. رتبه‌بندی در این سامانه بر اساس قوانین فازی که در جدول (۱) نمایش داده شده است، انجام می‌پذیرد.

(جدول-۱): مجموعه قوانین سامانه فازی در رتبه‌دهی اسناد
(Table-1): Rules for the main fuzzy system

شماره	برچسب Title	برچسب Body	لغات هم‌معنی	نقش دستوری	رتبه
1	High	High	NO	Same	VR[2]
2	High	High	NO	NotSame	VR
3	High	High	YES	Same	HR[3]
4	High	High	YES	NotSame	HR
5	High	Medium	NO	Same	VR
6	High	Medium	NO	NotSame	HR
7	High	Medium	YES	Same	HR
8	High	Medium	YES	NotSame	HR
9	High	Low	NO	Same	HR
10	High	Low	NO	NotSame	HR
11	High	Low	YES	Same	MR[4]
12	High	Low	YES	NotSame	MR
13	Medium	High	NO	Same	HR
14	Medium	High	NO	NotSame	MR
15	Medium	High	YES	Same	MR
16	Medium	High	YES	NotSame	MR
17-20	Medium	Medium	__	__	MR
21-24	Medium	Low	__	__	LR[5]
25	Low	Medium	NO	Same	MR
26	Low	Medium	NO	NotSame	MR
27	Low	Medium	YES	Same	LR
28	Low	Medium	YES	NotSame	LR
29-32	Low	Medium	__	__	NR[6]
33-36	Low	Low	__	__	NR

[2] Very Related
[3] High Related
[4] Medium Related
[5] Low Related
[6] Not Related

[DOI: 10.29252/jsdp.17.1.29]



(شکل-۷): تابع عضویت برای ارتباط میان اسناد

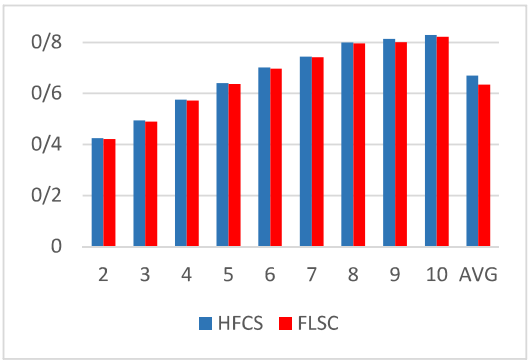
(Figure-7): Membership function for the linguistic variable Relevance

۸-۴- خوشه‌بندی در الگوریتم HFCS

در گام نهایی این سامانه و ارائه نتایج معنایی برای جستجو، از روش خوشه‌بندی سلسله‌مراتبی به همراه از مؤلفه‌های نامبرده استفاده شده است. برای این منظور از کتابخانهٔ «SharpCluster» و کدهای تولیدشده در نرم‌افزار «AggloCluster» نیز بهره‌گیری شده است. در این مرحله خروجی سامانه فازی که در مرحلهٔ قبل توضیح داده شد، به‌عنوان ورودی الگوریتم خوشه‌بندی منظور می‌شود. با اجرای این الگوریتم یک ساختار سلسله‌مراتبی به‌ازای تمام سندهای بازیابی‌شده تولید می‌شود و دو سندی که بیشترین شباهت به یکدیگر داشته باشند در یک ردیف از این سلسله‌مراتب قرار می‌گیرند. اسناد قرارگرفته در یک ردیف تنها وابسته به لغات نیستند و شباهت میان آن‌ها شباهت معنایی است (چه پنهان و چه آشکار). حال با توجه به این سلسله‌مراتب در هر مرحله از جستجو می‌توان سندهای مشابه جستجوی کاربر را به او پیشنهاد کرد، حتی سندهایی که هیچ‌گونه ارتباط لغوی با یکدیگر ندارند و تنها ارتباط آن‌ها معنایی است.

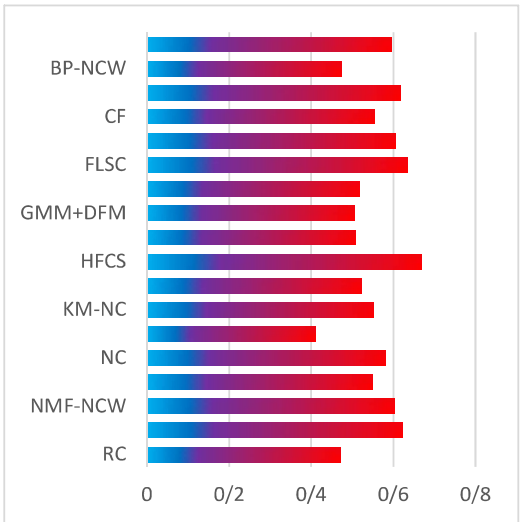
۵- تحلیل داده‌ها و آزمون فرضیات

در این روش به‌جای یک فهرست مسطح از نتایج جستجو یک ساختار درختی از نتایج جستجو ارائه شده است. درخت معنایی به این مفهوم است که سلسله‌مراتب در آن، ریشهٔ پرس‌وجوی کاربر است. در این درخت به‌جز ریشه، همه مفاهیم در گره‌های دیگر می‌تواند به‌عنوان جستجوی پیشرفته در نظر گرفته شود. مهم‌ترین مؤلفه بعد از دقت جستجو در آزمایش‌های ما، کشف ارتباط معنایی میان اصطلاحات موجود در جستجوی مورد نظر است که برای اندازه‌گیری این مؤلفه از مجموعه دادهٔ رویترز (REUTERS-21578, DATASET) استفاده شده که شامل ۲۱۵۷۸ سند و ۱۳۵ موضوع مجزاست. برای محاسبهٔ مقدار این ارتباط از روش کروسکال والیس H[1] استفاده شده است. در این آزمایش میانگین پاسخ به جستجو در الگوریتم ارائه‌شده (HFCS)، با استفاده از دو تا ده اصطلاح مرتبط اندازه‌گیری شده که تعداد این اصطلاحات متناظر با مقدار K در نظر گرفته شده است. این آزمایش به‌ازای هر مقدار K، پنجاه‌بار تکرار شده و سپس میانگین آن محاسبه شده است. برای اثبات نتایج این روش، HFCS با هفده روش موجود در مقاله [1] مقایسه شده که مقادیر این آزمایش در جدول (۲) و مقایسه میانگین آن‌ها در شکل (۹) نمایش داده شده است.



(شکل-۸): نمودار مقایسه دو روش HFCS و FLSC در کشف روابط معنایی

(Figure-8): Comparison of two methods of HFCS and FLSC in the discovery of semantic relations



(شکل-۹): نمودار مقایسه میانگین اسناد مرتبط در میان خوشه‌ها

(Figure-9): Comparison of the average of related documents among clusters

[1] Kruskal-Wallis H Test

[DOI: 10.29252/jsdp.17.1.29]

(جدول-۲): مقایسه ارتباط میان اسناد در روش HFCS با دیگر روش‌ها

(Table-2): Normalized mutual information comparision of HFCS method with other methods on "REUTERS-21578" dataset

K	2	3	4	5	6	7	8	9	10	Average
HFCS	0.425	0.494	0.575	0.641	0.701	0.745	0.799	0.814	0.829	0.669
FLSC	0.421	0.489	0.572	0.636	0.696	0.742	0.796	0.801	0.822	0.634
OBFDC	0.535	0.543	0.562	0.557	0.623	0.655	0.672	0.697	0.676	0.622
CCF	0.569	0.563	0.607	0.620	0.605	0.624	0.633	0.647	0.676	0.616
CF-NCW	0.496	0.505	0.595	0.616	0.644	0.615	0.660	0.660	0.665	0.606
NMF-NCW	0.494	0.500	0.586	0.615	0.637	0.613	0.654	0.659	0.658	0.602
AFC	0.546	0.521	0.567	0.562	0.613	0.659	0.663	0.687	0.667	0.596
NC	0.484	0.461	0.555	0.592	0.617	0.594	0.640	0.634	0.643	0.580
CF	0.480	0.429	0.503	0.563	0.592	0.556	0.613	0.609	0.629	0.553
NMF	0.480	0.426	0.498	0.559	0.591	0.552	0.603	0.601	0.623	0.548
KM	0.404	0.402	0.461	0.525	0.561	0.548	0.583	0.597	0.618	0.522
KM-NC	0.438	0.462	0.525	0.554	0.592	0.577	0.594	0.607	0.618	0.552
GMM	0.475	0.468	0.462	0.516	0.551	0.522	0.551	0.557	0.548	0.517
H2D-FCM	0.480	0.468	0.463	0.455	0.522	0.548	0.525	0.562	0.547	0.507
GMM+DFM	0.470	0.466	0.450	0.513	0.531	0.506	0.535	0.535	0.536	0.505
BP-NCW	0.391	0.377	0.431	0.478	0.493	0.500	0.519	0.529	0.532	0.472
RC	0.417	0.381	0.505	0.460	0.485	0.456	0.548	0.484	0.495	0.470
NB	0.466	0.348	0.401	0.405	0.409	0.404	0.435	0.411	0.418	0.411

همان‌گونه که در شکل (۸) مشخص است، روابط معنایی میان دو روش HFCS و FLSC مورد مقایسه قرارگرفته و هر بار کشف این روابط بر اساس تعداد واژگان موجود در جستجوی کاربر اندازه‌گیری شده که تعداد این واژه‌ها در شکل متناظر با متغیر K (مقادیر افقی) نمایش داده شده است. ستون سمت چپ در شکل (۸) بیان‌گر میزان کشف روابط میان اسناد در روش HFCS و ستون سمت راست بیان‌گر میزان کشف روابط میان اسناد در روش FLSC است؛ و همان‌گونه که انتظار می‌رفت در ابتدا با واژه‌های کمتر روش HFCS پاسخ بسیار نزدیک نسبت به روش FLSC خواهد داشت؛ اما با بالارفتن تعداد واژه‌ها روش FLSC به‌دلیل بزرگی ماتریس استفاده‌شده در روش FLSC و داشتن مراحل هنجارسازی بیشتر روش HFCS پاسخ بهتری بازمی‌گرداند؛ همچنین با توجه به میانگین این حالت‌ها روش HFCS در کشف روابط معنایی بهتر عملکرده و کارآمدتر خواهد بود. در ابتدا بیان کردیم که مهم‌ترین هدف از ارائه روش HFCS با بالابردن دقت در نتیجهٔ جستجو است. در این بخش نیازمند روشی برای ارزیابی نتیجهٔ آزمایش هستیم. زمانی که بتوان داده‌ها را به دو گروه مثبت و منفی تقسیم کرد، دقت نتایج یک آزمایش با استفاده از شاخص‌های حساسیت[1] و ویژگی[2] قابل‌اندازه‌گیری و توصیف است.

[1] Sensitivity
[2] Specificity

مؤلفه‌های دقت[3] موارد بازیابی‌شده از نتایج یک آزمایش را موردبررسی قرار می‌دهند، درحالی‌که مؤلفه‌های جامعیت[4] موارد بازیابی‌شده را از نتایج یک آزمایش را به همراه موارد صحیح بازیابی‌نشده موردبررسی قرار می‌دهند به همین دلیل است که این مؤلفه‌ها به‌طورمعمول به‌عنوان مؤلفه‌های حساسیت شناخته می‌شود. درنهایت هردو این موارد، به درک اندازه‌گیری ارتباط میان نتایج کمک می‌کند. دقت می‌تواند به‌عنوان یک اندازه‌گیری از نوع کیفیت دیده شود، درحالی‌که جامعیت یک اندازه‌گیری از نوع مقدار است. به عبارت ساده‌تر، می‌توان گفت دقت، بدان معنی است که پاسخ الگوریتم ما چقدر صحیح است، درحالی‌که جامعیت، به این معنی است که الگوریتم ما، چه مقدار از جواب صحیح را در نظر گرفته است [22]. باید توجه داشت که میزان صحیح‌بودن نتایج نیازمند واحدی است که بتواند این داوری را انجام دهد. برای این کار از اصطلاحی به نام رجوع به خبرگان[5] یا قضاوت کارشناسی استفاده می‌شود که به‌طور خاص به یک روش اشاره دارد که در آن قضاوت بر اساس یک مجموعه خاص از معیارها و یا تخصص انجام می‌پذیرد. این معیارها در یک پایگاه دانش خاص قرار می‌گیرد تا در صورت لزوم مورد استفاده قرار گیرند. در جدول (۳) نمونه‌ای از قوانین نمایش داده شده است.

[3] Precision
[4] Recall
[5] Expert Judgment

[DOI: 10.29252/jsdp.17.1.29]

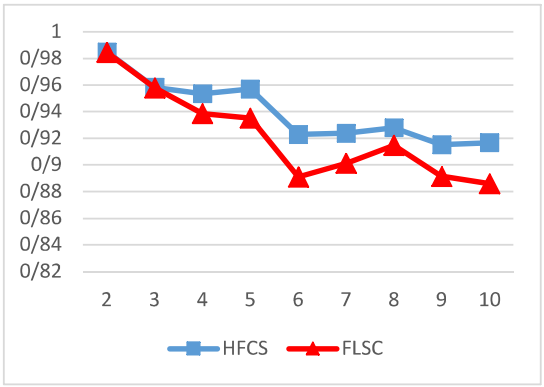
(جدول-۳): جدول حالت‌های قابل وقوع برای یک موضوع

(Table-3): The Contingency table for topic zi

Topic zi		Clustering Results	
		YES	NO
Expert Judgment	YES	Tpi	Fni
	NO	Fpi	Tni

در مجموعه‌داده رویترز ۱۳۵ موضوع مجزا وجود دارد که داوری بر روی نتایج بازیابی شده است بر اساس تعلق یک سند به آن موضوع انجام می‌پذیرد. در جدول (۳) مقدار Zi اشاره به موضوع مورد نظر دارد. در هنگام داوری بر روی هر سند چهار حالت ممکن است رخ دهد. ۱. پاسخ مثبت صحیح[1] (TP): سندی که به‌درستی مرتبط به موضوع مورد نظر تشخیص داده شود. ۲. پاسخ مثبت کاذب[2] (FP): سندی که به اشتباه مرتبط به موضوع مورد نظر تشخیص داده شود. ۳. پاسخ منفی صحیح[3] (TN): سندی که به‌درستی نامرتبط به موضوع مورد نظر تشخیص داده شود، نظر تشخیص داده شود. ۴. پاسخ منفی کاذب[4] (FN): سندی که به‌اشتباه نامرتبط به موضوع مورد نظر تشخیص داده شود. با توجه به مفاهیم نامبرده میزان دقت از طریق رابطهٔ زیر محاسبه خواهد شد و نتایج آن در شکل (۱۰) نمایش داده شده است.

$$Precision_i = \frac{TP_i}{TP_i + FP_i} \qquad (۴)$$



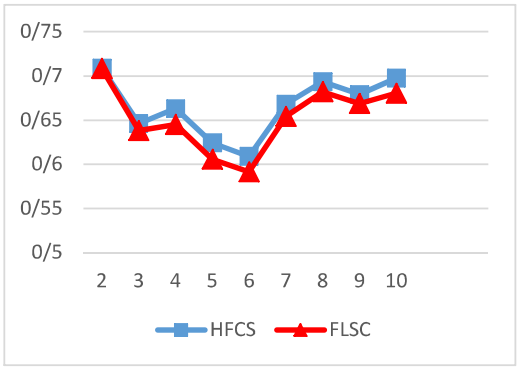
(شکل-۱۰): نمودار مقایسه دقت در دو روش HFCS و FLSC

(Figure-10): Comparison of precision in both HFCS and FLSC methods

در شکل (۱۰) میزان دقت دو روش HFCS و FLSC بر اساس رابطهٔ (۴) محاسبه شده است. همان‌طور که انتظار می‌رفت با افزایش واژگان جستجو که در این شکل متناظر با مقدار ستون عمودی است، میزان دقت در روش HFCS به‌مراتب بالاتر است. همچنین با بالا رفتن تعداد این واژه‌ها روش HFCS دقت دارای افت کمتری نسبت به روش FLSC است. پس از محاسبهٔ دقت نوبت به محاسبهٔ میزان جامعیت یک پاسخ خواهد رسید که از طریق رابطهٔ زیر محاسبه خواهد شد.

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \qquad (۵)$$

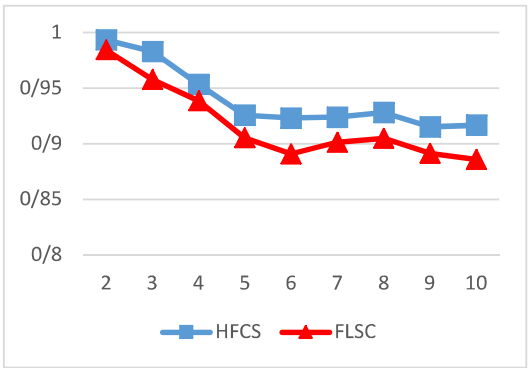
تفاوت میان این دو مقدار استفاده از پاسخ منفی کاذب به جای پاسخ مثبت کاذب است. با توجه به شکل (۱۱) مشاهده می‌شود که میزان جامعیت در این دو روش بسیار نزدیک بوده اما این میزان در روش HFCS با اندکی بهبود همراه بوده و روش ما علاوه‌بر افزایش دقت، پاسخ جامع‌تری ارائه می‌دهد.



(شکل-۱۱): نمودار مقایسه جامعیت در دو روش HFCS و FLSC

(Figure-11): Comparison of recall in both HFCS and FLSC methods

در زمینهٔ علوم و مهندسی، صحت[5] یک سامانه اندازه‌گیری، درجه نزدیک‌بودن اندازه‌گیری یک مقدار به مقدار واقعی آن مقدار است.



(شکل-۱۲): نمودار مقایسه صحت در دو روش HFCS و FLSC

(Figure-12): Comparison of Accuracy test in both methods of HFCS and FLSC

[1] True Positive
[2] False Positive
[3] True Negative
[4] False Negative
[5] Accuracy

[DOI: 10.29252/jsdp.17.1.29]

$$Accuracy_i = \frac{TP_i+TN_i}{TP_i+TN_i+F_i+FN_i} \quad (۶)$$

میزان صحت یک الگوریتم از طریق رابطهٔ (۶) محاسبه می‌شود. همان‌طور که در انتهای بخش (۱) توضیح داده شد، به‌دلیل استفاده از منطق فازی و خوشه‌بندی سلسله‌مراتبی، دانش عمیق‌تری از معانی لغات استفاده‌شده در اسناد وب مشتق شده و شباهت میان اسناد قابل اعتمادتر است؛ به همین دلیل است که مقادیر روش پیشنهادی به مقادیر واقعی نزدیک‌ترند و شکل (۱۲) به‌طورکامل گویای این مسأله است. به‌طورکلی، با توجه به مقادیر به‌دست‌آمده از اندازه‌گیری‌های مکرر از یک مجموعه، مجموعه‌ای را می‌توان دقیق نامید که مقادیر آن به یکدیگر نزدیک باشند؛ این در حالی است که مجموعه‌ای را می‌توان صحیح نامید که متوسط مقادیر در اندازه‌گیری‌های مکرر نزدیک به مقادیر واقعی باشند؛ در این حالت می‌توان نتیجه گرفت که این دو مفهوم مستقل از یکدیگر هستند و به یک مجموعه جواب می‌تواند ترکیب‌های مختلفی از این دو مقدار اطلاق شود.

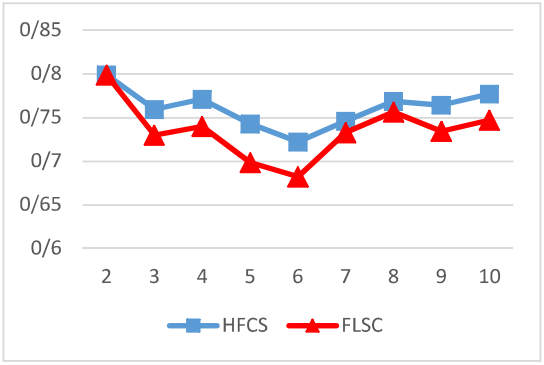
آزمون F از سری آزمون‌های تجزیه و تحلیل واریانس (ANOVA[1]) است. این آزمون در علوم رایانه، به‌طور خاص در بازیابی اطلاعات، یادگیری ماشین، میانگین هم‌ساز[2]، دقت و جامعیت مورد استفاده قرار می‌گیرد. این آزمون اغلب یک نمره از عملکرد کلی سامانه برای ارزیابی الگوریتم‌ها و زیرسامانه‌های مختلف آن ارائه می‌دهد. این نام‌گذاری توسط *جورج دبلیو اسندکور*[3] به‌افتخار *سر رونالد فیشر*[4] انجام شد. فیشر در سال ۱۹۲۰ نسبت واریانس را توسعه داد. این آزمون برای مقایسه میانگین و هم‌قوارگی چند جامعه استفاده می‌شود [23]. درواقع F نسبت دو واریانس است؛ یعنی آزمون F یک رابطه میان پاسخ مثبت صحیح و مجموعهٔ انتخاب‌شده توسط سامانه، نسبت به پاسخ مثبت صحیح و پاسخ مثبت حقیقی[5] ایجاد می‌کند که این پاسخ مثبت حقیقی با متغیر β نمایش داده می‌شود و طبق رابطهٔ زیر محاسبه می‌شود.

$$F_{\beta}=\frac{(\beta^2+1)\times Precision_i \times Recall_i}{\beta^2 \times Precision_i + Recall_i} \quad (۷)$$

نمرهٔ آزمون F می‌تواند به‌عنوان میانگین وزنی دقت و جامعیت، در نظر گرفته شود که در آن نمرهٔ یک در بهترین ارزش خود و نمرهٔ صفر در بدترین ارزش خود تفسیر می‌شود

[1] Analysis of Variance
[2] Harmonic
[3] George W. Snedecor
[4] Sir Ronald A. Fisher
[5] Real Positive

[24]. درنهایت روش HFCS را با روش FLSC از لحاظ میزان دقت و جامعیت و آزمون مقایسه کردیم که نتایج این مقایسه در شکل (۱۳) نمایش داده شده و همان‌طور که این شکل به‌طورکامل گویاست روش HFCS دارای برتری نسبی از بیشتر جهات نسبت به FLSC است.



(شکل-۱۳): نمودار مقایسه آزمون F در دو روش HFCS و FLSC
(Figure-13): Comparison of F test in both methods of HFCS and FLSC

۶- نتیجه‌گیری و کارهای آینده

همان‌طور که گفته شد یکی از مشکلات اساسی کاربران وب بحث جستجو است، همچنین کاربران وب بیشتر علاقه‌مند به کشف مفاهیم اطلاعات در وب هستند تا استفادهٔ مستقیم از آن‌ها درواقع کاربران از موتورهای جستجو توقع درک متن موردنظر خود را دارند که این درک موجب به‌وجودآمدن نسل جدیدی از وب‌ها با نام وب‌معنایی شده است. در وب‌معنایی علاوه‌بر بازیابی نتایج جستجو اسناد مرتبط با جستجو که شاید ارتباط مستقیم یا ارتباط لغوی با موضوع نداشته باشند نیز بر اساس هوش مصنوعی طراحی‌شده به کاربر پیشنهاد می‌شوند. در این جستجو روابط معنایی پنهان نیز در نظر گرفته می‌شوند تا به نتیجهٔ بهتری در جستجو دست یابیم. باید توجه داشت که چندمعنایی‌ها و عبارات وابسته، محدود به فناوری جستجو هستند. یک اصطلاح مفرد، همیشه قادر به شناسایی یک مفهوم نهفته در یک سند نیست، برای مثال: اصطلاح «شبکه» در ارتباط با واژهٔ «رایانه»، «ترافیک» یا «عصبی» نشان‌دهنده مفاهیم مختلف است؛ اما یک گروه از واژگان می‌تواند یک مفهوم را به‌صورت واضح نمایش دهند. برای شناسایی و تبعیض موضوعات صحیح در یک مجموعه از اسناد، ترکیب ویژگی‌های و روابط و احتمالات از اهمیت خاصی برخوردار است. به این منظور همهٔ ویژگی‌های موجود در اسناد بازیابی‌شده با روش‌های نامبرده توسط سامانه فازی یکسان‌سازی شده و به‌ازای هر

[DOI: 10.29252/jsdp.17.1.29]

جستجو از طرف کاربر یک ساختار سلسله‌مراتبی معنایی با استفاده از وزن فازی و خوشه‌بندی سلسله‌مراتبی تشکیل شده است که هر سطح آن درستی مربوط به مباحث در مجموعه‌ای از اسناد را مشخص می‌کند. ما به‌طور مؤثر مشاهده کردیم که چنین راه‌کاری با منطق فازی و استفاده از خوشه‌بندی سلسله‌مراتبی از اسناد وب، بیشینه ارتباط را در میان روش‌های موجود خواهد داشت. بر اساس آزمایش‌های ما، درمی‌یابیم که روش HFCS یک راه بسیار خوب برای سازمان‌دهی اطلاعات بدون ساختار و نیمه‌ساختاری است که داده‌ها را به چند موضوع معنایی تقسیم می‌کند و خود یک مدل مؤثر برای خوشه‌بندی اسناد وب به‌صورت خودکار است.

۱-۶- پیشنهادهای آینده

موارد زیر را به‌عنوان پیشنهاد جهت پروژه‌های آینده می‌توان مطرح کرد:

۱. در این مقاله از تعدادی واژه‌نامهٔ ایستا برای تشخیص واژگان مترادف، نقش دستوری و ایست‌واژه‌ها استفاده شده است. تبدیل هریک از واژه‌نامه‌ها به واژه‌نامه‌های پویا می‌تواند گامی مؤثر در پیشبرد این روش باشد.

۲. استفاده از ضریب دیریکله یا روش‌های دیگر برای هنجارسازی داده‌ها به جای زیر سامانه فازی

۳. ارائه روش‌های جهت افزایش سرعت جستجو

۷- مراجع 7-References

[1] I. J. Chiang, C. C. H. Liu, Y. H. Tsai, and A. Kumar, "Discovering Latent Semantics in Web Documents Using Fuzzy Clustering," *IEEE Transactions on Fuzzy Systems,* vol. 23, no. 6, pp. 2122-2134, 2015.

[2] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, 2009, pp. 496.

[3] K. R. Pole and V. R. Mote, "Name Entity Recognition and Natural Language Processing for Improvised Fuzzy clustering in Web Documents," *Intenational Journal of Advance Research in Science and Engineering,* vol. 6, no. 09, 2017.

[4] D. Bollegala, Y. Matsuo, and M. Ishizuka, "A Web Search Engine-Based Approach to Measure Semantic Similarity between Words," *IEEE Transactions on Knowledge and Data Engineering,* vol. 23, no. 7, pp. 977-990, 2011.

[5] B. Jiang, Z. Li, H. Chen, and A. G. Cohn, "Latent Topic Text Representation Learning on Statistical Manifolds," *IEEE Transactions on Neural Networks and Learning Systems,* pp. 1-12, 2018.

[6] C. S. S. Kumar, M. Mohanapriya, and C. Kalaiarasan, "A new approach for information retrieval in semantic web mining involving weighted relationship," in *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, 2017, pp. 1-4.

[7] R. Zhao and K. Mao, "Fuzzy Bag-of-Words Model for Document Representation," *IEEE Transactions on Fuzzy Systems,* vol. 26, no. 2, pp. 794-804, 2018.

[8] M. K. Rafsanjani, Z. A. Varzaneh, and N. E. Chukanlo, "A Survey Of Hierarchical Clustering Algorithms," *Journal of Mathematics and Computer Science(JMCS),* vol. 5, no. 3, pp. 229-240, 2012.

[9] H. Park, K. Kwon, A. i. Z. Khiati, J. Lee, and I. J. Chung, "Agglomerative Hierarchical Clustering for Information Retrieval Using Latent Semantic Index," in *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*, 2015, pp. 426-431.

[10] D. Rahmawati, G. A. P. Saptawati, and Y. Widyani, "Document clustering using sequential pattern (SP): Maximal frequent sequences (MFS) as SP representation," in *2015 International Conference on Data and Software Engineering (ICoDSE)*, 2015, pp. 98-102.

[11] G. Bordogna and G. Pasi, "Hierarchical-Hyperspherical Divisive Fuzzy C-Means (H2D-FCM) Clustering for Information Retrieval," in *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, 2009, vol. 1, pp. 614-621.

[12] Nisha and P. J. Kaur, "Cluster quality based performance evaluation of hierarchical clustering method," in *2015 1st International Conference on Next Generation Computing Technologies (NGCT)*, 2015, pp. 649-653.

[13] C. Subbalakshmi, G. R. Krishna, S. K. M. Rao, and P. V. Rao, "A Method to Find Optimum Number of Clusters Based on Fuzzy Silhouette on Dynamic Data Set," *Procedia Computer Science,* vol. 46, pp. 346-353, 2015.

[14] M. Kaur and U. Kaur, "Comparison between k-means and hierarchical algorithm using query redirection," *International Journal of Advanced Research in Computer Science and Software Engineering,* vol. 3, no. 7, 2013.

[15] J. T. Chien, "Hierarchical Theme and Topic Modeling," *IEEE Transactions on Neural Networks and Learning Systems,* vol. 27, no. 3, pp. 565-578, 2016.

[DOI: 10.29252/jsdp.17.1.29]

[16] Q. Mao, W. Zheng, L. Wang, Y. Cai, V. Mai, and Y. Sun, "Parallel Hierarchical Clustering in Linearithmic Time for Large-Scale Sequence Analysis," in *2015 IEEE International Conference on Data Mining*, 2015, pp. 310-319.

[17] A. J. C. Trappey, C. V. Trappey, F. C. Hsu, and D. W. Hsiao, "A Fuzzy Ontological Knowledge Document Clustering Methodology," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics),* vol. 39, no. 3, pp. 806-814, 2009.

[18] T. X. Society, S. Wang, Q. Jiang, and J. Z. Huang, "A Novel Variable-order Markov Model for Clustering Categorical Sequences," *IEEE Transactions on Knowledge and Data Engineering,* vol. 26, no. 10, pp. 2339-2353, 2014.

[19] A. Ensan and Y. Biletskiy, "Matching semi-structured documents using similarity of regions through fuzzy rule-based system," in *Industrial Conference on Data Mining*, 2013, pp. 205-217: Springer.

[20] D. A. Grossman and O. Frieder, *Information retrieval: Algorithms and heuristics*. Springer Science & Business Media, 2012.

[21] A. N. Langville and C. D. Meyer, *Google's PageRank and beyond: The science of search engine rankings*, Princeton University Press, 2011.

[22] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation," *Journal of Machine Learning Technologies,* pp. 37–63, 2011.

[23] A. Dalli, "Adaptation of the f-measure to cluster based lexicon quality," *EACL 2003 Workshop on Evaluation,* pp. 51–56, 2003.

[24] C. J. v. RIJSBERGEN, *INFORMATION RETRIEVAL*. Newton, MA: Butterworth, 1979.

حمید شکرزاده تحصیلات خود را در مقطع کارشناسی در رشته مهندسی کامپیوتر گرایش سخت‌افزار در دانشگاه آزاد اسلامی واحد تهران مرکز به پایان رسانید و مدرک کارشناسی ارشد خود را در رشته مهندسی فناوری اطلاعات گرایش شبکه‌های کامپیوتری از دانشگاه آزاد اسلامی واحد قزوین اخذ کرده است. ایشان مدرک دکترای خود را در رشته مهندسی کامپیوتر گرایش معماری کامپیوتر از دانشگاه آزاد اسلامی واحد علوم تحقیقات و فناوری اخذ و پس از آن به عضویت هیئت علمی دانشگاه آزاد واحد پردیس درآمدند. علایق پژوهشی ایشان شامل شبکه‌های حس‌گر بی‌سیم، الگوریتم مسیریابی، شبیه‌سازی شبکه، پردازش زبان طبیعی، داده‌کاوی و پردازش اسناد وب‌معنایی است.

نشانی رایانامه ایشان عبارست از:
shokrzadeh@gmail.com

بهنام طاهری خامنه دانشجوی مقطع کارشناسی ارشد رشته مهندسی فناوری اطلاعات گرایش شبکه‌های رایانه‌ای و زمینه‌های فعالیت وی یادگیری ماشین، داده‌کاوی و معماری وب است.

نشانی رایانامه ایشان عبارست از:
behnam.taheri@pardisiau.ac.ir

[DOI: 10.29252/jsdp.17.1.29]