



# خوشه‌بندی سلسله‌مراتبی فازی برای کشف روابط معنایی پنهان در اسناد وب معنایی

بهنام طاهری خامنه و حمید شکرزاده\*

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

## چکیده

رشد انبوه اطلاعات در وب مشکلاتی را به دنبال 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 مسئله مهم‌ترین مانع در سرویس‌دهی بهتر به کاربران وب است. هدف این مقاله ارائه 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 مرتبه با آن، مورد بررسی قرار می‌گیرد. پاسخ به دست آمده 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 به دست آمده آزمون F اعمال می‌شود که اغلب به عنوان 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 ارائه شده در این مقاله می‌تواند پاسخ دقیق‌تر و جامع‌تری نسبت به روش‌های مشابه خود ارائه دهد و به طور میانگین دقت را تا ۰/۲۲ 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 Engineering, 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

\*نویسنده عهده‌دار مکاتبات

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

سال ۱۳۹۹ شماره ۱ پیاپی ۴۳

**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

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 یافته شده نداشته باشد. این امر نیازمند به دست آوردن شباهت معنایی میان استناد است. اندازه‌گیری دقیق شباهت معنایی بین واژه‌های موجود در استناد، یک مشکل مهم در وب کاوی، بازیابی اطلاعات و پردازش زبان طبیعی است. برنامه‌های کاربردی وب کاوی نیاز به توانایی دقیق اندازه‌گیری شباهت معنایی بین مفاهیم یا موجودیت‌ها دارند [4]. روش‌های شباهت معنایی با استفاده از تعییه لغت به عنوان یکی از عناصر تشکیل‌دهنده بردار شباهت، هر متن را به عنوان یک بردار میانگین از لغات، نمایش می‌دهند [5]. کمار<sup>۴</sup> و همکاران [6] یک رویکرد جدید را بر اساس شباهت کسینوسی<sup>۵</sup> ارائه کرده‌اند. در این رویکرد ساختار معنایی مبتنی بر شباهت کسینوسی نه تنها 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 هنجارسازی‌شده‌ای از وقوع یک اصطلاح پایه در سند اشاره دارد. برای شمارش تعداد وقوع بک اصطلاح پایه، مدل کیف واژگان<sup>۶</sup>، تطبیق دقیق واژگان را انجام می‌دهد که می‌تواند به عنوان نگاشت سخت از واژگان، به اصطلاح پایه محسوب شود. همچنین این روش‌ها توانایی ضبط معنای در پشت داده‌های متداول را ندارند. برای حل این مشکل ژئو<sup>۷</sup> 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 رشته لغت استفاده می‌شود. ورودی پایه برای معنای وب معنایی ورودی کاربر است، اما دقت خروجی

<sup>4</sup> Kumar

<sup>5</sup> Cosine Similarity

<sup>6</sup> Normalized

<sup>7</sup> Bag of Words

<sup>8</sup> Zhao

## ۱- مقدمه

موتورهای جستجو، ابزاری ضروری برای یافتن و استخراج اطلاعات مورد نظر از صفحات وب است که در جمع‌آوری و پالایش<sup>۹</sup> اطلاعات به کاربران کمک می‌کند [1]. در طول دهه گذشته، بهینه‌سازی بی‌وقفه در روش‌های بازیابی اطلاعات، موتورهای جستجو را به سطح جدیدی از کیفیت رهنمون کرده است؛ به طوری که جستجوی وب به یک منبع استاندارد و اغلب ارجح برای یافتن اطلاعات تبدیل شده است [2]. استناد 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 افعالات بالا بین اصطلاحات در استناد گاهی نشان‌گر معنای مبهم است. این ابهام نه تنها برای کاربر بلکه برای موتورهای جستجو 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 محتوا آن نمی‌توان اظهار کرد. ساختار نمایشی وب به گونه‌ای 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<sup>۱۰</sup> 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 استفاده مستقیم از آن‌ها، درواقع کاربران از موتورهای جستجو توقع درک متن موردنظر خود را دارند؛ بنابراین ما به دنبال بازنمایی از اطلاعات هستیم که در آن ماشین با سرزدن به یک تارنما معنای هر یک از بخش‌های<sup>۱۱</sup> موجود در آن را درک کند. وب معنایی به دنبال راهی می‌گردد تا کاربر پرس‌وجوهی مورد نیازش را به مرورگر

<sup>1</sup> Filter

<sup>2</sup> Index

<sup>3</sup> Items



فراهم کند؛ علاوه‌بر موارد بالا، روش پیشنهادی برای کاربر، به صورت سلسله‌مراتبی استفاده از مشابه‌ترین اسناد و ب را از لحاظ معنایی فراهم می‌کند. با استفاده از این دو روش، دانش عمیق‌تری از معانی لغات استفاده شده در اسناد و ب مشتق شده و شباهت میان اسناد قابل اعتمادتر است. در روش HFCS به جای گروه‌بندی اسناد به فهرست‌های گسترده‌ای از نتایج، اسناد به خوشه‌های قابل فهم برای ماشین تبدیل می‌شوند که این امر در راستای غایت نهایی و ب 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 نحوه پیاده‌سازی روش پیشنهادی بیان می‌شود. در بخش بعد نتایج و تحلیل خروجی آن‌ها مورد بررسی می‌گیرد و نحوه مقایسه 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 معرفی شده‌اند 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 به صورت درختی ارائه می‌شود که هر سطح آن مراحل خوشه‌بندی را نمایش می‌دهد. به این درخت سلسله‌مراتبی دندوگرام<sup>۵</sup> می‌گویند. روش‌های خوشه‌بندی بر اساس ساختار سلسله‌مراتبی تولیدی توسط آن‌ها به‌طور معمول به دو دستهٔ زیر تقسیم می‌شوند:

۱. روش تجمعی؟ این روش به روش پایین به بالا نیز معروف است. در روش تجمعی، ابتدا هر سند به عنوان یک خوشه در نظر گرفته می‌شود. در ادامه با به کارگیری یک الگوریتم تکرارشونده، در هر مرحله خوشه‌های دارای ویژگی‌های مشابه با یکدیگر ادغام شده و خوشهٔ جدیدی را تشکیل می‌دهند. این روند تا شکل‌گیری یک ساختار سلسله‌مراتبی از خوشه‌ها ادامه می‌یابد.

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 اشیاء 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 پیشنهادی می‌شود. در گام نخست داده‌های مرتبط با درخواست ارسالی کاربر با بهره‌گیری از الگوریتم (TF-IDF) وزن‌دهی می‌شوند. محاسبه شباهت ابتدایی میان اسناد با استفاده از شباهت کسینوسی انجام می‌پذیرد. برای کشف روابط معنایی پنهان و همچنین بالابردن دقت جستجو مؤلفه‌هایی<sup>۱</sup> نظری نقش دستوری<sup>۲</sup> و لغات مترادف به الگوریتم جستجو اضافه می‌شود. علاوه‌بر مؤلفه‌های نامبرده، مؤلفه‌های معنایی که به عنوان مهم‌ترین بخش در تشخیص روابط معنایی در نظر گرفته می‌شوند، استخراج خواهد شد. این مؤلفه‌ها به عنوان فراداده<sup>۳</sup> در مجموعه داده‌های ما قرار گرفته‌اند. در گام بعدی مجموع این مؤلفه‌ها برای وزن‌دهی به عنوان ورودی یک زیرسامانه فاری ارسال می‌شوند؛ درنهایت با استفاده از سامانه تصمیم‌گیری فازی و استفاده از خوشه‌بندی سلسله‌مراتبی الگوریتمی با نام HFCS<sup>4</sup> معرفی می‌کنیم تا در صدد حل مقداری از مشکلات نامبرده برآییم.

در سال‌های اخیر تلاش‌های بسیاری برای خوشه‌بندی داده‌های حجمی و کشف روابط معنایی میان آن‌ها صورت گرفته است. بر اساس بررسی‌های انجام‌شده در عمل روش مشابه دیگری وجود ندارد که با استفاده از خوشه‌بندی تجمعی و تصمیم‌گیری فازی امکان ارائه بهترین پاسخ به جستجوی کاربر را بر روی داده‌های وب‌معنایی

<sup>1</sup> Parameters

<sup>2</sup> Part of Speach

<sup>3</sup> MetaData

<sup>4</sup> Hierarchical Fuzzy Clustering Semantics



۲. روش تفکیکی<sup>۱</sup>: از این روش به عنوان روش بالا به پایین نیز یاد می‌شود. در این روش ابتدا تمام اسناد به عنوان یک خوشه در نظر گرفته می‌شود؛ سپس با به کارگیری یک الگوریتم تکرارشونده، در هر مرحله داده‌ای که کمترین شباهت را با داده‌های دیگر دارد، به عنوان 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<sup>۲</sup> به همراه خوشه‌بندی سلسله‌مراتبی با رویکرد تجمعی، پیشنهاد شده است. مشاهده می‌شود که پاسخ‌های جستجو با روش‌های معمول خوشه‌بندی نتایج<sup>۳</sup>، راضی‌کننده نیستند و لغات می‌توانند به صورت بهتری خوشه‌بندی شوند. هنگامی که روش‌های اولیه خوشه‌بندی نتایج از مدل مجموعه‌ای از واژگان استفاده می‌کنند، پیداکردن روابط معنایی به صورت ساده میسر نخواهد بود. هدف از ترکیب این دو روش، پیداکردن مترادف‌ها و چندمعنایی‌های است؛ همچنین این روش از نظر کیفیت خوشه‌ای بسیار قابل توجه 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<sup>۴</sup> K برای خوشه‌بندی داده‌های متوالی استفاده کرده و ویژگی‌های اسناد را به عنوان 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 ارائه شکل مناسب از ویژگی‌های سند ضروری است. برای این منظور دو نوع الگوی ترتیبی وجود دارد: توالی تکرار یک لغت<sup>۵</sup> و بیشینه توالی تکرار<sup>۶</sup> 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 مجموعه داده‌ها توسط تعداد مؤلفه‌ها در خوشه‌های مورد هدف مشخص می‌شود.

<sup>1</sup> Divisive

<sup>2</sup> Latent Semantic Indexing

<sup>3</sup> Search Result Clustering

<sup>4</sup> K-Means

<sup>5</sup> Frequent Word Sequence

<sup>6</sup> Maximal Frequent Sequence

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 به صورت 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 اندازه یک بی‌نظمی<sup>۷</sup> تعیین می‌کند و تصمیم می‌گیرد که آیا یک گره می‌تواند بر اساس تراکم<sup>۸</sup> تقسیم شود.

در مقاله [12] تعدادی از روش‌های خوشه‌بندی مختلف مورد تجزیه و خوشه‌بندی تفکیکی. در این مقاله خوشه‌بندی تجمعی و خوشه‌بندی تجزیه از سه برای تجزیه و تحلیل اندازه‌گیری کیفیت خوشه‌ها از سه مؤلفه: اندازه‌گیری سازگاری، گذشت زمان و شاخص سیلوئت<sup>۹</sup> استفاده شده است. روش سیلوئت یکی از متداول‌ترین و بهترین روش‌های اعتبارسنجی الگوریتم‌های خوشه‌بندی است. این روش کیفیت خوشه را به صورت 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، به عنوان واژگان و جملات در اسناد متنی، یک روش یادگیری ساختاری برای استخراج موضوعات و مضماین<sup>۱۰</sup> پنهان، میان جملات و واژگان ارائه شده است. در این روش رابطه بین مضماین و موضوعات، تحت گروه‌بندی‌های مختلف، از طریق یک روش بدون نظرارت، بدون محدود کردن تعداد خوشه‌ها مورد بررسی قرار گرفته است. در این روش از فرآیند شکستن چوب<sup>۱۱</sup> برای ارائه شباهت میان جملات

<sup>7</sup> Entropy

<sup>8</sup> Density

<sup>9</sup> Silhouette

<sup>10</sup> Themes

<sup>11</sup> Stick-Breaking



وزن‌دهی فازی است، یک الگوریتم خوشه‌بندی سلسه‌مراتبی جدید با وزن فازی ارائه می‌شود.

در مقاله [1] برای کشف معانی پنهان در اسناد وب، یک الگوریتم بر اساس فضای توپولوژیکی تشکیل شده از متغیرهای زبانی فازی ارائه شده است. الگوریتم پیشنهادی ویژگی‌ها را از اسناد وب با استفاده از روش فضای تصادفی شرطی<sup>۶</sup> 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 هفته<sup>۷</sup> استفاده شده است. این روش با بهره‌گیری از خوشه‌بندی فازی وابستگی و هماهنگی ویژگی‌های هر مجموعه، سلسه‌مراتبی از مجتمعهای معنایی متصل به نام مفاهیم را به وجود 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 هزینه 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 هستی‌شناسی پویا برای استخراج ویژگی‌ها استفاده کردند. علاوه بر این 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 هنگارسازی انجام گیرد؛ به همین دلیل در روش HFCS عامل شباهت را به شباهت میان عبارت موردنظر کاربر با اسناد موجود در مجموعه داده تغییر دادیم در این حالت تنها نیازمند مقایسه سند با عبارت موردنظر کاربر خواهیم بود. در این حالت هنگارسازی را بهازای هر سند به دو مرحله کاهش می‌دهیم.

مختلف استفاده شده است. در این مقاله یک مدل موضوعی سلسه‌مراتبی ایجاد می‌شود که با استفاده از الگوریتم‌های فاقد مؤلفه بیزین<sup>۱</sup>، 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 سلسه‌مراتبی درخت معنایی برای انتخاب جملات نمایش داده می‌شود.

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

در مقاله [17] یک رویکرد جدید خوشه‌بندی سلسه‌مراتبی برای خودسازماندهی اسناد علمی، ارائه شده 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. علاوه بر این، یک روش کنترلی با استفاده از منطق فازی برای مطابقت خوشه‌های مستقل و هستی‌شناسی مشتق شده از آن‌ها مورداستفاده قرار گرفته است. نتایج نشان می‌دهد که رویکرد خوشه‌بندی فازی مبتنی بر هستی‌شناسی از رویکرد K-Means در دقت، جامعیت، آزمون F و بی‌نظمی شانون<sup>۲</sup> برتر است.

در مقاله [18] یک چارچوب<sup>۳</sup> نوین از متغیرهای مارکوف به نام توزیع احتمالی شرطی وزن‌دار<sup>۴</sup> 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<sup>۵</sup> نامیده شده است؛ درنهایت بر اساس یک چارچوب بهینه‌سازی آبشراری که ترکیبی از مدل‌های توزیع احتمالی شرطی وزن‌دار و شاخص

<sup>1</sup> Bayesian<sup>2</sup> Shannon<sup>3</sup> Framework<sup>4</sup> Weighted Conditional Probability Distribution<sup>5</sup> Weighted Fuzzy Indicator

### ۳- منطق فازی

منطق فازی یک نوع منطق است که روش‌های متعدد نتیجه‌گیری در مغز بشر را جایگزین الگوهای ساده‌تر ماشینی می‌کند. درواقع می‌توان این طور استدلال کرد که مغز بشر به ورودی‌های اطلاعاتی دقیق نیازی ندارد، بلکه قادر است تا کنترل تطبیقی را به صورت بالایی انجام دهد و این در مورد ماشین نیز صادق است. ساده‌ترین تلقی برای تعریف منطق فازی این است که منطق فازی جواب یک سوال را به جای تقسیم به دو بخش درست یا نادرست، در اصل به یک محدوده جواب، در این‌بین 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 گستردگر است و می‌توان با استفاده از قواعد منطق فازی، جواب‌های فازی مناسب با پرسش را ارائه کرد.

منطق فازی می‌کوشد، دانستنی‌هایی که ناحدودی قابل توصیف و بیان زبان‌شناختی بوده، ولی امکان کمی کردن آن‌ها با کمک ریاضیات سنتی به طور معمول وجود ندارد را به صورتی 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 در واقع، یک عبارت زبان طبیعی است که به یک مقدار کمیت خاص اشاره دارد و در اصطلاح مانند مترجم عمل می‌کند و به کمک تابع عضویت، نشان داده می‌شود؛ تابع عضویت به شما اجازه می‌دهد تا یک واژه زبانی را کمی‌سازی کرده و یک مجموعه فازی را به صورت نموداری نمایش دهید.

اگرچه سامانه‌های فازی پدیده‌های غیرقطعی و نامشخص را توصیف می‌کند با این حال نظریه فازی یک نظریه دقیق است.

### ۴- ارائه روش پیشنهادی

همان‌طور که در بخش‌های قبل بیان شد خوش‌بندی فرایندی است که در نتیجه آن گروهی از خوش‌ها بر اساس معیار شباht ایجاد می‌شوند؛ بنابراین معیارهای شباht نقش مهمی را در خوش‌بندی ایفا می‌کنند. یکی از روش‌های رایج برای ارزیابی شباht بین الگوها استفاده از معیار فاصله است. بیشترین معیار فاصله موردن استفاده در الگوریتم‌های مختلف استفاده از فاصله اقلیدسی است. فاصله اقلیدسی

فصلنامه



به‌طور معمول برای خوش‌بندی 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 استناد به منظور بالا بردن دقت جستجو در HFCS<sup>۱</sup> 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 به تنها یکی از 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 استناد 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<sup>۲</sup> است؛ که هدف از آن ارائه استناد به عنوان 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 نهفته‌ای احتمالی<sup>۳</sup> را نام برد که برخلاف سایر روش‌های فضایی برداری مجموعه‌ای از مفاهیم مرتبط با استناد و اصطلاحات آن را تولید می‌کنند. ایده اصلی این روش‌ها رخدادن روابط پنهان، میان اصطلاحات نزدیک‌تر در استناد مشابه، ممکن است. پژوهش‌گران روش‌های استخراج اطلاعات را به دو روش: تحت نظرارت<sup>۴</sup> و بدون نظرارت<sup>۵</sup> تقسیم‌بندی کرده‌اند. هر دو روش استخراج ویژگی تحت نظرارت و بدون نظرارت از هر اصطلاح به عنوان ابعادی از یک مدل فضایی برای نشان‌دادن یک سند در یک واحد خاص استفاده می‌کنند. یکی از شایع‌ترین راه‌های بدون نظرارت در

<sup>1</sup> Hierarchical Fuzzy Clustering Semantics

<sup>2</sup> Vector Space Model

<sup>3</sup> Probabilistic Latent Semantic Analysis

<sup>4</sup> Supervised

<sup>5</sup> Unsupervised



جستجوشده از طرف کاربر تحت فرآیند تحلیل واژگانی<sup>۴</sup> به لغات جدا تبدیل می‌شود برای مثال: زمانی که کاربر عبارت: «کتاب صدسال تنها بی از مارکز» را جستجو می‌کند، فهرستی از لغات به صورت [کتاب، صد، سال، تنها بی، از، مارکز] تشکیل شده و به ازای هر یک از این لغات در صورت تکرار نبودن تمام اسناد موجود جستجو خواهد شد. پس از انجام جستجو متوجه می‌شویم که نتیجه جستجو مطلوب نخواهد بود؛ زیرا در همین مثال لغت "از" در بسیاری از متنون فارسی موجود است و زمانی که جستجو انجام می‌شود تمام اسناد شامل آن بازیابی خواهد شد. به این گونه واژگان ایستوازه<sup>۵</sup> گفته می‌شود. ایستوازه‌ها لغاتی هستند که برخلاف تکرار فراوان در متن، از لحاظ معنایی دارای اهمیت پایینی هستند، مانند "اگر"، "و"، "ولی"، "که" و غیره. در نگاه اولیه واژگان ربط و تعریف، ایستوازه به نظر می‌آیند؛ اما سیاری از افعال، افعال کمکی، اسم‌ها، قیدها و صفات نیز ایستوازه شناخته شده‌اند. در اغلب کاربردهای متن، حذف این واژگان، نتایج پردازش را بهشت ببود می‌دهد و سبب کاهش بار محاسبات و افزایش سرعت خواهد شد. به همین دلیل این واژگان اغلب در مرحله پیش‌پردازش، حذف می‌شوند. برای زبان فارسی چندین فهرست از این واژگان منتشر شده است که به‌طور میانگین شامل پانصد واژه است. پس از حذف ایستوازه‌ها، در مرحله بعد اقدام به ریشه‌یابی<sup>۶</sup> لغات می‌کنیم.

ریشه‌یابی عبارت است از به‌دست‌آوردن ریشه واژگان با حذف پسوندها و پیشوندها به‌طوری که واژگان با ریشه یکسان دارای شکل یکسان شوند. معمول‌ترین هدف ریشه‌یابی استفاده از پایه و اساس یک کلمه، برای شناخت واژگان مشابه است. در سامانه بازیابی اطلاعات استفاده از واژگان ریشه‌یابی شده 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 محاسبه شباهت، اختصاص مؤلفه‌های افزایش دقت و اختصاص وزن در سامانه فازی است که در ادامه به تشریح هریک خواهیم پرداخت.

وزن دهی مدارک روش (TF-IDF) است؛ که در این مقاله از این روش به‌منظور وزن دهی اولیه اسناد استفاده شده است. تلاش‌های بسیاری برای اندازه‌گیری شباهت بین اسناد مبتنی بر مدل فضایی برداری مانند شباهت کسینوسی، شباهت اقلیدسی<sup>۱</sup>، ضربی جاکارد<sup>۲</sup> و دیگر روش‌ها انجام شده است. در روش پیشنهادی ما، برای محاسبه شباهت اولیه میان اسناد از ایده بردار اصطلاحات به همراه روش شباهت کسینوسی استفاده شده است. ساختار تولیدشده بر اساس معماری وب‌معنایی در لایه توصیف منابع<sup>۳</sup> از راه فراداده افزوده شده 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. به ازای هر مؤلفه افزوده شده به سامانه، محاسبه شباهت میان اسناد انجام گرفته و مجموع این شباهت‌ها برخلاف دیگر روش‌های موجود به عنوان ورودی سامانه فازی قرار خواهد گرفت. پس از وزن دهی سامانه فازی، خروجی سامانه جستجو با استفاده از وزن شباهت اولیه و وزن خروجی از سامانه فازی با استفاده از خوشه‌بندی سلسله‌مراتبی رتبه‌بندی خواهد شد و بر اساس این رتبه خروجی جستجو به ترتیب نمایش داده شده و اسناد مرتبط از لحاظ معنایی به کاربر پیشنهاد می‌شود. در بخش بعدی به توضیح دقیق‌تر و مoshkafanه راجع به مراحل نامیرده خواهیم پرداخت.

#### ۱-۴- پیاده‌سازی روش HFCS

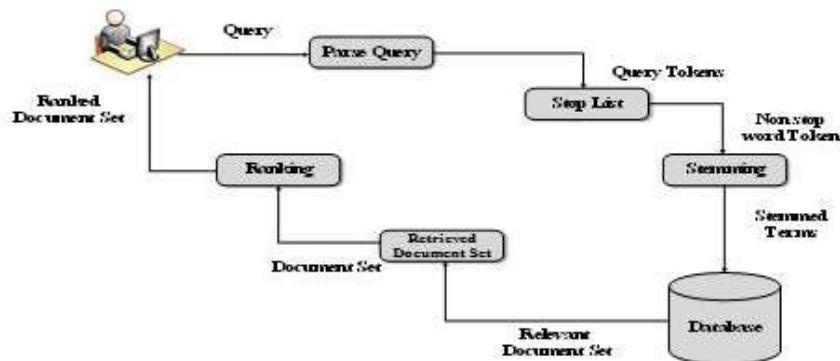
همان‌طور که مشاهده می‌شود در شکل (۱) مراحل کلی جستجو و رتبه‌دهی اسناد تا بازگشت پاسخ به کاربر نمایش داده شده است.

در گام نخست جستجوی معمول هنگامی که کاربر جستجوی موردنظر خود را ارسال می‌کند، عبارت

<sup>1</sup> Euclidean Similarity

<sup>2</sup> Jaccard Coefficient

<sup>3</sup> Resource Description Framework (RDF)

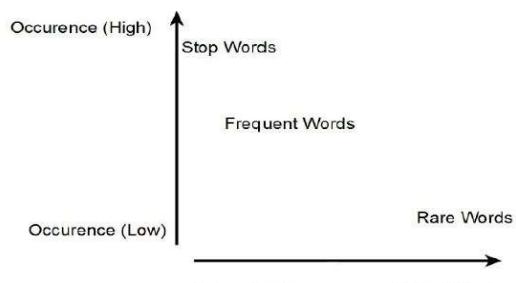


(شکل-۱): فلوچارت جستجو در اسناد  
(Figure-1): Search System FlowChart

سند) و IDF نشان‌دهنده بسامد معکوس یک واژه در سند است (تعداد تکرار یک واژه در کل اسناد). وزن TF-IDF از طریق رابطه زیر محاسبه می‌شود.

$$W_{TF-IDF} = (1 + \log TF) \times \log \left( \frac{N}{df_i} \right) \quad (2)$$

در اینجا N نشان‌دهنده کل اسناد موجود و  $df_i$  نشان‌دهنده فهرست اسناد شامل واژه مورد نظر است.



[۲۰] شکل-۲: نمودار اهمیت لغات در روش TF-IDF  
(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 زاویه 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 زاویه میان آن‌هاست که در رابطه  $(3)$  نمایش داده شده است.

$$\text{Similarity}(A, B) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}} \quad (3)$$

#### ۴-۲- محاسبه شباهت

همان‌طور که گفته شد، با افزایش تعداد و ارگان در جستجو نیازمند سازوکاری برای طبقه‌بندی اسناد و نمایش بهترین نتیجه به کاربر هستیم، مدل فضای برداری یک چارچوب ساختاریافته ریاضی با مجموعه‌ای از عناصر<sup>۱</sup> است که مجموعه این عناصر بردار نامیده می‌شوند و هر بردار می‌تواند به‌وسیله 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 به صورت مجموعه‌ای از عناصر نمایش داده شود.

$$V = a_1 v_1 + a_2 v_2 + \dots + a_k v_n \quad (1)$$

در اینجا  $v_n$  نشان‌دهنده عناصر و  $a_k$  برابر وزن آن‌هاست. برای محاسبه 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 مسئله 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 به طول طبیعی بیشتر از دیگران تکرار خواهند شد، اگرچه چندان اهمیتی در معنی نداشته باشند. ارزش TF-IDF از دو تابع<sup>۲</sup> TF<sup>۳</sup> و IDF به دست می‌آید که در آن TF نشان‌دهنده بسامد یک واژه در یک سند (تعداد تکرار یک واژه در یک

<sup>1</sup> Elements

<sup>2</sup> Term Frequency and weighting

<sup>3</sup> Inverse Document Frequency

بهزای هر مؤلفه‌ای که برای بالابدن 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 تعییش شده است. با این روش دیگر محدودیتی در افزایش تعداد مؤلفه‌ها نخواهیم داشت و تنها بهزای هر مؤلفه نیازمند اضافه کردن قوانین جدیدی هستیم. همچنین با داشتن چنین زیرسامانه‌ای دیگر برای استفاده از مؤلفه‌های مختلف نیازمند هنجارسازی نخواهیم بود؛ چون تمامی مقادیر توسط سامانه 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 کتابخانه «AI FuzzyLogicLibrary» که معروف به «FuzzyLogicLibrary» 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
4: IR ← FuzzyInputVariable (inputTitle, 0.0, 1.0);
5: for (i = 0; i < Count (IT); i++) do
6:   IR.Add (IT, FuzzyType (ITH1, ITH2, ..., ITHn))
     <IR.Add ("inputTerm",
TriangularMembershipFunction(0.39, 0.5, 0.61))
7: end for
8: OR ← FuzzyOutputVariable (outputTitle, 0.0, 1.0);
9: for (j = 0; j < Count (OT); j++) do
10:   OR.Add (T, FuzzyType (OTH1, OTH2, ..., OTHm))
11: end for
12: 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 مقایسه 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

در رابطه (۳): مقادیر A و B دو سند هستند که قرار است شباهت میان آن‌ها سنجیده شود و مقدار  $\alpha$  نشان‌دهنده ویژگی استخراج شده 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.

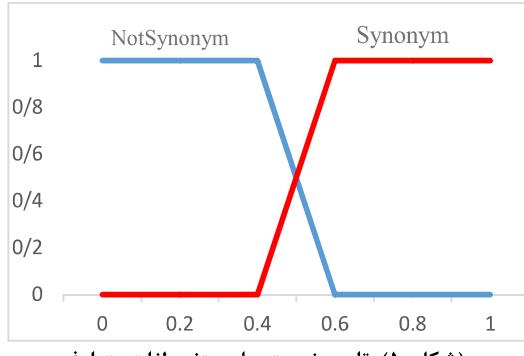
#### ۴-۴- وزن دهنی بر اساس نقش دستوری

با پردازش هر جمله می‌توان نقش دستوری یک لغت در آن جمله را مشخص کرد و بهزای آن یک وزن به آن لغت اختصاص داد. عملیات مشخص کردن نقش بر اساس یک سری برچسب‌های استانداردشده انجام می‌ذیرد. استفاده از این نوع وزن دهنی 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 به عنوان اسم مورداستفاده قرار گیرد به معنای کتاب و اگر به عنوان فعل مورد استفاده قرار گیرد به معنای رزرو کردن است. در بیشتر سامانه‌های جستجو از جمله 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 استفاده معنایی از این برچسب‌ها را از دست خواهیم داد و برچسب‌هایی که برچسبی غیر از نقش «اسم» دارند با ارزش یکسان تلقی خواهند شد.علاوه بر این اگر از روش بالا برای نقش دستوری TF-IDF استفاده شود امکان تأثیر مؤلفه‌های متعدد در وزن میسر نخواهد بود. در روش پیشنهادی، برچسب‌های معنایی تنها در دو گروه قرار نگرفته‌اند. در این روش به جای شمارش تمام برچسب‌ها از برابری نقش دستوری استفاده شده است. برای این منظور ابتدا نیازمند یک لغتنامه معنایی شامل برچسب‌های دستوری استانداردشده و همچنین سامانه‌ای برای برچسب‌زنی اسناد خواهیم بود؛ برای این منظور از کتابخانه «POSTagger» برای برچسب‌زنی استفاده شده است. بعد از به‌دست آوردن نقش دستوری نیازمند روشهای وزن دهنی نقش دستوری به هریک از اسناد هستیم. برای محاسبه این وزن یک زیرسامانه<sup>۱</sup> فازی طراحی شده است که در بخش بعدی به توضیح درباره آن خواهیم پرداخت.

<sup>1</sup> Sub System



«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. درصورتی که جستجو به‌ازای تمام این لغات انجام گیرد، نتیجه 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 بازیابی شده 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 این که 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 مؤلفه نقش دستوری به‌عنوان ورودی به زیرسامانه فازی تحویل داده شده وتابع عضویت فازی به‌ازای مؤلفه 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- وزن دهی بر اساس برچسب‌های معنایی

در این بخش برای رتبه‌بندی میان اسناد بازیابی شده 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 محاسبه شباهت کسینوسی به‌ازای هر سند، شباهت کسینوسی به‌ازای هر برچسب معنایی محاسبه خواهد شد.

تکرار واژگان یا وزن TF می‌توان در صدد 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 مؤلفه نقش دستوری یا هر مؤلفه 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
1: FSR [ ] ← Null           <FuzzySubSystemRules;
2: SSInput [ ] ← Null        <SubSystemInput
3: SSOOutput [ ] ← Null      <SubSystemOutput
4: FIV ← Null               <FuzzyInputVariable (Title, min, max)
5: FOV ← Null               <FuzzyOutputVariable (Title, min, max)
6: Result [ ] ← Null
7: if FSR = Null then
8:     FSR ← LoadFuzzySubSystemRules( );
    ▷ (see DefindFuzzySubSystemRules(Table...))
9: else
10:    for (j = 1; j < Count (PSCV); j++) do
    ▷ Count = 5
11:       FIV ← FSR.SelectInputByTitle (PSCV[j]
12:       FOV ← FR.SelectOutputByTitle
    (outputTitle);
13:       SSInput.add (FIV, PSC[j]);
14:       SSOOutput ← CalculateFuzzyWeight
    (SSInput);
15:       Result [j-1] ← SSOOutput [FOV]
16:    end for
16: end if
17: Return Result

```

(شکل-۴): شبکه 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<sup>۱</sup> در ایجاد فراداده معرفی شده‌اند. این استاندارد به‌عنوان سامانه ساده ساختاردهی دانش در وبمعنایی معرفی شد. این استاندارد علاوه‌بر نقش دستوری و لغات مترادف شامل مؤلفه‌هایی از قبیل طبقه‌بندی<sup>۲</sup>، رده‌بندی<sup>۳</sup>، سرموضع<sup>۴</sup> (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 کتابخانه

<sup>1</sup> Simple Knowledge Organization System

<sup>2</sup> Classification

<sup>3</sup> Taxonomies

<sup>4</sup> Subject-Heading





- <sup>2</sup> Very Related  
<sup>3</sup> High Related  
<sup>4</sup> Medium Related  
<sup>5</sup> Low Related  
<sup>6</sup> Not Related

در بخش‌های قبل مؤلفه‌های افزایش دقت به عنوان ورودی زیرسامانه فازی در نظر گرفته شد و به ازای آن وزن فازی متناظر با آن محاسبه شد. در این مرحله برای استفاده از این سامانه فازی تمام مؤلفه‌های محاسبه شده به عنوان 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 هم معنی	نقش دستوری	رتبه
1	High	High	NO	Same	VR <sup>2</sup>
2	High	High	NO	NotSame	VR
3	High	High	YES	Same	HR <sup>3</sup>
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 <sup>4</sup>
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 <sup>5</sup>
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 <sup>6</sup>
33-36	Low	Low	—	—	NR

هریک از این برچسبها می‌تواند به عنوان مؤلفه‌ای در افزایش دقت جستجو ایفای نقش کند. همچنین با این روش می‌توان میان برچسبها اولویت ایجاد کرد، برای مثال زمانی که نتیجه جستجو در عنوان یک سند یافت شد (در برچسب Title) اولویت بیشتری نسبت به یافتن نتیجه در متن خبر (در برچسب Body) داشته باشد یا می‌توان به واژگان کلیدی در انتهای هر سند اولویت بیشتری اختصاص داد. علاوه بر امکانات نامبرده با این روش می‌توان به ازای رده‌بندی مردمی<sup>۱</sup> یا دیگر رتبه‌بندی‌ها نیز یک برچسب اختصاص داد و مقدار به دست آمده را در نتیجه جستجو اعمال کرد؛ درنهایت برای یکسان‌سازی داده‌ها و تبدیل آن به وزن فازی، میزان شباهت به دست آمده از هر برچسب معنایی، به عنوان مؤلفه 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 به ازای هر برچسب معنایی محاسبه شد؛ اما بر اساس مشکلات نامبرده امکان استفاده از تمام این برچسبها به صورت یکجا وجود نداشت. برای این منظور یک سامانه فازی طراحی شده است تا عمل رتبه‌دهی و تجمعیت تمام این مؤلفه‌ها و درنهایت تصمیم‌گیری برای نمایش بهترین نتیجه جستجو به کاربر را انجام دهد.

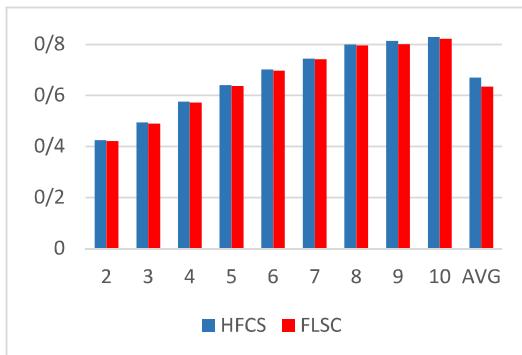
<b>Name:</b> Hierarchical Fuzzy Clustering Semantics (HFCS)
<b>Input:</b> Primary Similarity Scores [ ] PSC, Primary Similarity Scores Variable [DocumnetNumber, Title, Body, Part of Speech, Synonym] PSCV, Cosine Similarity CS, Number of Documents ND
<b>Output:</b> Hierarchical Cluster
1: FSSW [ ] ← Null ↳ FuzzySubSystemWeight
2: FW ← Null ↳ FuzzyWeight
3: Clusters[ ] ← Null
4: LoadFuzzyRules(); ↳ Load Rules From Table-1
5: <b>for</b> (i = 0; i < ND; i++) <b>do</b>
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: <b>end</b>
10: NC ← Number of Cluster
11: <b>while</b> NC != 1 <b>do</b>
12:     MergeTwoClosestCluster (Clusters);
13:     NC ← NC - 1
14: <b>end while</b>

(شکل-۶): شبکه‌کد الگوریتم HFCS

(Figure-6): Pseudo Code of the HFCS Algorithm

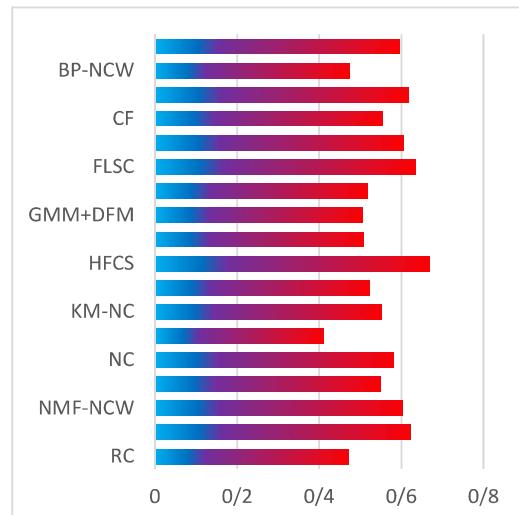
<sup>1</sup> Taxonomy

21578، DATASET) استفاده شده که شامل ۲۱۵۷۸ سند و ۱۳۵ موضوع مجاز است. برای محاسبه مقدار این ارتباط از روش کروسکال والیس <sup>۱</sup> استفاده شده است. در این آزمایش میانگین پاسخ به جستجو در الگوریتم ارائه شده (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، پنجاه بار تکرار شده و سپس میانگین آن محاسبه شده است. برای اثبات نتایج این روش، HFCS با هفده روش موجود در مقاله [۱] 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

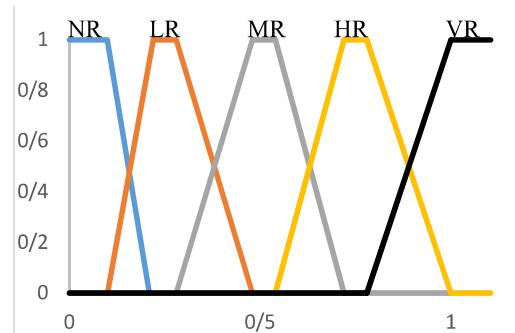
(Figure-8): Comparison of two methods of HFCS and FLSC in the discovery of semantic relations



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

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

<sup>۱</sup> Kruskal-Wallis H Test



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

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

#### ۴-۴- خوشبندی در الگوریتم

در گام نهایی این سامانه و ارائه نتایج معنایی برای جستجو، از روش خوشبندی سلسله‌مراتبی به همراه از مؤلفه‌های نامبرده استفاده شده است. برای این منظور از کتابخانه «SharpCluster» و کدهای تولیدشده در نرم‌افزار «AggloCluster» نیز بهره‌گیری شده است. در این مرحله خروجی سامانه فازی که در مرحله قبل توضیح داده شد، به عنوان ورودی الگوریتم خوشبندی منظور می‌شود. با اجرای این الگوریتم یک ساختار سلسله‌مراتبی به‌ازای تمام سندهای بازیابی شده تولید می‌شود و دو سندی که بیشترین شباهت به یکدیگر داشته باشند در یک ردیف 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 به جای 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، ریشه پرس‌وحشی کاربر است. در این درخت به جزء ریشه، همه مفاهیم در گره‌های دیگر می‌تواند به عنوان 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 داده Reuters- (REUTERS-

<sup>3</sup> Precision<sup>4</sup> Recall<sup>5</sup> Expert Judgment

(جدول-۲): مقایسه ارتباط میان اسناد در روشن 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
OBFDCC	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

مؤلفه‌های دقت<sup>۳</sup> موارد بازیابی شده از نتایج یک آزمایش را موردنرسی قرار می‌دهند، درحالی که مؤلفه‌های جامعیت<sup>۴</sup> موارد بازیابی شده را از نتایج یک آزمایش را به همراه موارد صحیح بازیابی نشده موردنرسی قرار می‌دهند به همین دلیل است که این مؤلفه‌ها به طور معمول به عنوان مؤلفه‌های حساسیت شناخته می‌شود. درنهایت هردو این موارد، به درک اندازه‌گیری ارتباط میان نتایج کمک می‌کند. دقت عبارت ساده‌تر، می‌توان گفت دقت، بدان معنی است که پاسخ الگوریتم ما چقدر صحیح است، درحالی که جامعیت، به این معنی است که الگوریتم ما، چه مقدار از جواب صحیح را در نظر گرفته است [22]. باید توجه داشت که میزان صحیح بودن نتایج نیازمند واحدی است که بتواند این داوری را انجام دهد. برای این کار از اصطلاحی به نام رجوع به خبرگان<sup>۵</sup> یا قضاوت کارشناسی استفاده می‌شود که به طور 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.

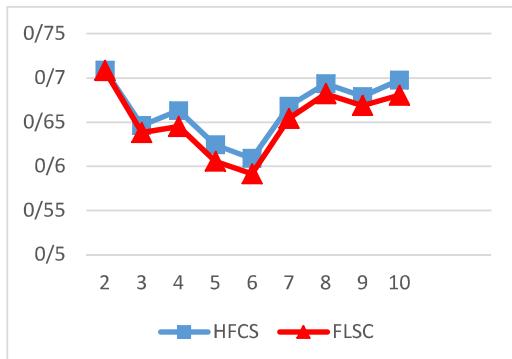
همان‌گونه که در شکل (۸) مشخص است، روابط معنایی میان دو روش HFCS و FLSC مورد مقایسه قرار گرفته و هر بار کشف این روابط بر اساس تعداد واژگان موجود در جستجوی کاربر اندازه‌گیری شده که تعداد این واژه‌ها در شکل متناظر با متغیر K (مقادیر افقی) نمایش داده شده است. ستون سمت چپ در شکل (۸) بیان‌گر میزان کشف روابط 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 FLSC پاسخ بسیار نزدیک نسبت به روش HFCS خواهد داشت، اما با بالا رفتن تعداد واژه‌ها روش FLSC به دلیل بزرگی ماتریس استفاده شده در روش HFCS 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 بالا رفتن دقت در نتیجه جستجو است. در این بخش نیازمند روشی برای ارزیابی نتیجه آزمایش هستیم. زمانی که بتوان داده‌ها را به دو گروه مثبت و منفی تقسیم کرد، دقت نتایج یک آزمایش با استفاده از شاخص‌های حساسیت<sup>۱</sup> و ویژگی<sup>۲</sup> قابل اندازه‌گیری و توصیف است.

<sup>1</sup> Sensitivity<sup>2</sup> Specificity

مقدار ستون عمودی است، میزان دقت در روش HFCS به مرتب بالاتر است. همچنین با بالا رفتن تعداد این واژه‌ها روش HFCS دقت دارای افت کمتری نسبت به روش FLSC است. پس از محاسبه دقت نوبت به محاسبه میزان جامعیت یک پاسخ خواهد رسید که از طریق رابطه زیر محاسبه خواهد شد.

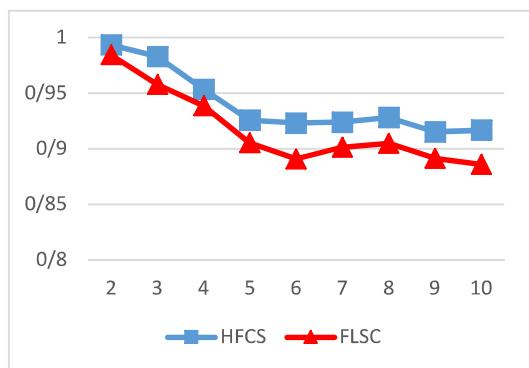
$$Recall_i = \frac{TP_i}{TP_i + FN_i} \quad (5)$$

تفاوت میان این دو مقدار استفاده از پاسخ منفی کاذب به جای پاسخ مثبت کاذب است. با توجه به شکل (11) 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 علاوه بر افزایش دقت، پاسخ جامعه‌تری ارائه می‌دهد.



(شکل-۱۱): نمودار مقایسه حامیت در دو روش HFCS و FLSC  
(Figure-11): Comparison of recall in both HFCS and FLSC methods

در زمینه علوم و مهندسی، صحت<sup>۵</sup> یک سامانه اندازه‌گیری، درجه نزدیک بودن 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

<sup>5</sup> Accuracy

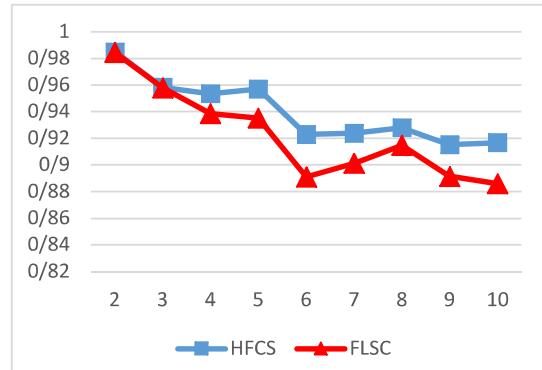
(جدول-۳): جدول حالت‌های قابل وقوع برای یک موضوع

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

Topic zi		Clustering Results	
		YES	NO
Expert Judgment	YES	T <sub>pi</sub>	F <sub>ni</sub>
	NO	F <sub>pi</sub>	T <sub>ni</sub>

در مجموع عدداده 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<sup>۱</sup> (TP): سندی که به درستی مرتبط به موضوع مورد نظر تشخیص داده شود. ۲. پاسخ مثبت کاذب<sup>۲</sup> (FP): سندی که به اشتباه مرتبط به موضوع مورد نظر تشخیص داده شود. ۳. پاسخ منفی صحیح<sup>۳</sup> (TN): سندی که به درستی نامرتبط به موضوع مورد نظر تشخیص داده شود، نظر تشخیص داده شود. ۴. پاسخ منفی کاذب<sup>۴</sup> (FN): سندی که به اشتباه نامرتبط به موضوع مورد نظر تشخیص داده شود. با توجه به مفاهیم نامبرده میزان دقت از طریق رابطه زیر محاسبه خواهد شد و نتایج آن در شکل (۱۰) نمایش داده شده است.

$$Precision_i = \frac{TP_i}{TP_i + FP_i} \quad (4)$$



(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 رابطه (۴) محاسبه شده است. همان‌طور که انتظار می‌رفت با افزایش واژگان جستجو که در این شکل متناظر با

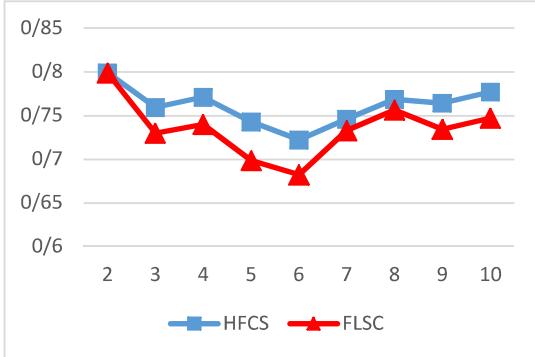
<sup>1</sup> True Positive

<sup>2</sup> False Positive

<sup>3</sup> True Negative

<sup>4</sup> False Negative

[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 به‌طور کامل 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

$$Accuracy_i = \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \quad (6)$$

میزان صحت یک الگوریتم از طریق رابطه (۶) محاسبه می‌شود. همان‌طور که در انتهای بخش (۱) توضیح داده شد، به‌دلیل استفاده از منطق فازی و خوشه‌بندی سلسه‌مراتبی، دانش عمیق‌تری از معانی لغات استفاده شده 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) به‌طور کامل گویای این مسئله است. به‌طور کلی، 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 می‌توانند ترکیب‌های مختلفی از این دو مقدار اطلاق شود.

آزمون F از سری آزمون‌های تجزیه و تحلیل واریانس (ANOVA<sup>۱</sup>) است. این آزمون در علوم رایانه، به‌طور خاص در بازیابی اطلاعات، یادگیری ماشین، میانگین همساز<sup>۲</sup>، 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 جرج دبلیو اسندرکور<sup>۳</sup> به افتخار سر رونالد فیشر<sup>۴</sup> 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 مجموعه انتخاب‌شده توسط سامانه، نسبت به پاسخ مثبت صحیح و پاسخ مثبت حقیقی<sup>۵</sup> ایجاد می‌کند که این پاسخ مثبت حقیقی با متغیر  $\beta$  نمایش داده می‌شود و طبق رابطه زیر محاسبه می‌شود.

$$F_{\beta} = \frac{(\beta^2 + 1) \times Precision_i \times Recall_i}{\beta^2 \times Precision_i + Recall_i} \quad (7)$$

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

<sup>1</sup> Analysis of Variance

<sup>2</sup> Harmonic

<sup>3</sup> George W. Snedecor

<sup>4</sup> Sir Ronald A. Fisher

<sup>5</sup> Real Positive

- Manifolds," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1-12, 2018.
- [6] C. S. S. Kumar, M. Mohanapriya, and C. Kalaiaras, "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. Saptaawati, and Y. Widjiani, "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 (H2DFCM) 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.

جستجو از طرف کاربر یک ساختار سلسله‌مراتبی معنایی با استفاده از وزن فازی و خوشبندی سلسله‌مراتبی تشکیل شده است که هر سطح آن درستی مربوط به مباحثت در مجموعه‌ای از اسناد را مشخص می‌کند. ما به طور مؤثر مشاهده کردیم که چنین راهکاری با منطق فازی و استفاده از خوشبندی سلسله‌مراتبی از اسناد وب، بیشینه ارتباط را در میان روش‌های موجود خواهد داشت. بر اساس آزمایش‌های ما، در می‌یابیم 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 خوشبندی اسناد وب به صورت خودکار است.

## ۶- پیشنهادهای آینده

- موارد زیر را به عنوان پیشنهاد جهت پژوهش‌های آینده می‌توان مطرح کرد:
۱. در این مقاله از تعدادی واژه‌نامه ایستا برای تشخیص واژگان متراffد، نقش دستوری و ایستو واژه‌ها استفاده شده است. تبدیل هریک از واژه‌نامه‌ها به واژه‌نامه‌های پویا می‌تواند گامی مؤثر در پیشیرد 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

## 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

شبیه‌سازی شبکه، پردازش زبان طبیعی، داده‌کاوی و پردازش استناد و بمعنایی است.  
نشانی رایانمای ایشان عبارت 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 رایانمای ایشان عبارت از:  
behnam.taheri@pardisiau.ac.ir

- [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 شکرزاوه** 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. علائق پژوهشی ایشان شامل شبکه‌های حس‌گر بی‌سیم، الگوریتم مسیریابی،



