Karafan

Quarterly Scientific Journal of Technical and Vocational University

Autumn 2023, Vol. 20, No. 3, p. 485-507

Journal Homepage: https://karafan.tvu.ac.ir/?lang=en

TVU

doi :10.48301/KSSA.2023.382583.2426

Implementation of a Noisy Hyperlink Removal System: Using the Semantic and Relational Approach of the DBpedia Ontology

Kazem Taghandiki[1*]

[1]Department of Computer Engineering, Technical and Vocational University (TVU), Tehran, Iran.

ARTICLE INFO

Article Type:
Original Research

Received: 01.27.2023
Revised: 05.06.2023
Accepted: 06.11.2023

Keyword:
Semantic Web
Noisy Hyperlinks
Ontology
Reasoner
Semantic Similarity
Relatedness Similarity

***Corresponding Author:**
Kazem Taghandiki
Email: *ktaghandiki@tvu.ac.ir*

ABSTRACT

With the rapid expansion and growth of web data, the web graph structure, which is a graphical representation of the web world, is getting larger and larger and has gradually changed from a content structure to a non-content structure. The presence of junk data such as noisy hyperlinks in the web structure graph has caused problems for many link mining algorithms and reduced the speed and efficiency of information retrieval algorithms. Research has been conducted to remove noisy hyperlinks using structural and string approaches. These approaches incorrectly remove some useful hyperlinks and are unable to detect noisy hyperlinks in some situations. In this paper, a dataset of noisy and useful hyperlinks was first created by an interactive crawler using website crawling. Then, through semantic web approaches and facilities such as the Dbpedia ontology, attention was paid to the semantic and relational structure of these hyperlinks. This was followed by activating the DBpedia ontology reasoner, the process of removing noisy hyperlinks from the web structure graph taking place. The tests performed on this system showed the accuracy and capability of Semantic Web technologies to remove noisy hyperlinks.

E-ISSN: 2538-4430
ISSN: 2382-9796

EXTENDED ABSTRACT

Introduction

In recent years, the ability to create an infinite number of web pages, together with the extremely large amounts of data generated in various fields of technology, has given rise to a challenging concept known as 'big data'. In 2021, nearly 49 billion web pages will be indexed by Google and Bing crawlers. Clearly, there is a significant increase in the number of web pages on the internet, which has led to the growth of the web structure graph. As a result, it is extremely difficult to navigate and explore the structure of the web due to spam data such as noisy hyperlinks, hence the need for a mechanism to eliminate spam hyperlinks. A number of studies have been conducted to detect and eliminate spam links from the structure of the web. However, the proposed methods are highly dependent on the string and structural characteristics of hyperlinks, while ignoring their semantic and relatedness structures. In this paper, the semantic and relatedness structures of hyperlinks at both the page and site levels were considered. Using semantic web technologies, such as ontologies and reasoners, noisy hyperlinks were removed.

First, a data set of hyperlinks was created in a separate process. Then, using semantic web technologies such as ontologies and reasoners, the concept of the hyperlinks on the source page and the concept of the target page were analyzed semantically and relationally. The analysis can be used to determine whether a hyperlink is noisy or useful.

The proposed system takes the constructed dataset as input; each row of the dataset is composed of the class mapped from the hyperlink context topic of the source page, the class mapped from the topic of the target page, a field indicating the noisy or useful nature of the hyperlink from the user's perspective, and the domain name of the source page. Then, noisy hyperlinks are detected and the results are compared with those of the user to determine the extent to which each semantic or relatedness property in the ontology contributes to correct detection. It is also possible to identify which queries lead the user to noisy hyperlinks and which domains contain the noisiest hyperlinks. The experiments demonstrate the accuracy, power and scalability of Semantic Web technologies in eliminating noisy hyperlinks.

Methodology

As shown in Figure 1, the semantic and relatedness system for eliminating noisy hyperlinks consisted of three general steps.

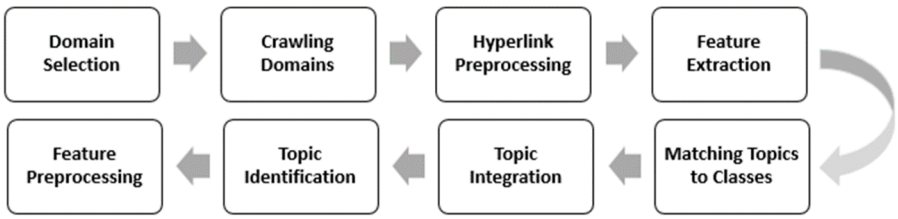


Figure 1. Implementation process of the proposed approach.

1- **Domain selection:** The main idea behind this domain selection approach is to crawl the domains that are retrieved by search engines in response to frequent queries.
2- **Crawling Domains:** the crawler starts to explore the domain and eventually a list of links in the domain is obtained.
3- **Hyperlink Preprocessing:** Incompatible hyperlinks must be removed in the preprocessing step.
4- **Feature Extraction:** In order to ensure that the topic of the surrounding text of the hyperlink as well as the topic of the target page are detected with sufficient accuracy, a number of features must be extracted from web pages.
5- **Feature Preprocessing:** This is able to perform the topic identification process with higher accuracy and lower error rate. In this paper, this was achieved by using typical text mining preprocessing techniques such as stop words, token normalization, case folding and stemming.
6- **Topic Identification:** The purpose of this stage is to determine the topic of the hyperlink text and the target page using the previously extracted features.
7- **Topic Integration:** The topics from the previous stage, located in multiple text files, are combined and integrated to create a single well-formed input for the matching stage.
8- Matching Topics to Classes: The purpose of this stage is to match the topics from the previous stage to the classes of the DBpedia ontology.

Results and discussion

Figure 2 visualizes the values of six commonly used performance measures in information retrieval systems. The extent to which the reasoner uses different semantic and relatedness properties in the DBpedia ontology to represent semantic and relatedness similarities between the subject and the object is shown in Figure 3. The percentages of hyperlink types from both perspectives are visualized in Figure 4 which shows that the proposed approach is able to distinguish between noisy and useful hyperlinks as accurately as an expert user.

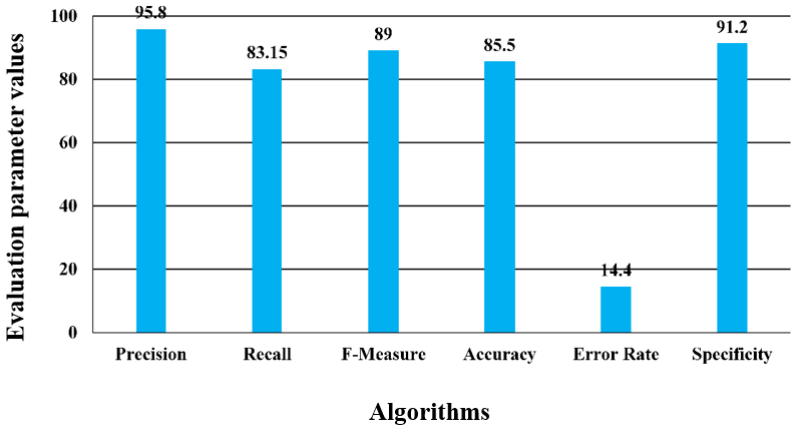


Figure 2. Performance measures of the proposed approach.

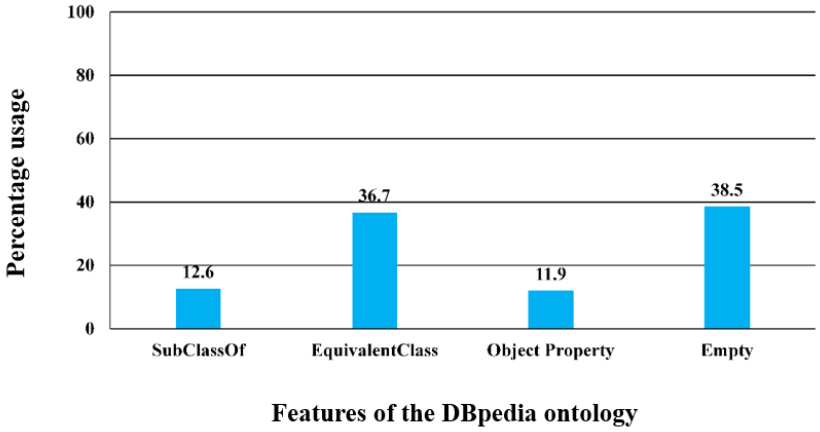


Figure 3. Percentage of DBpedia ontology properties used by reasoner.

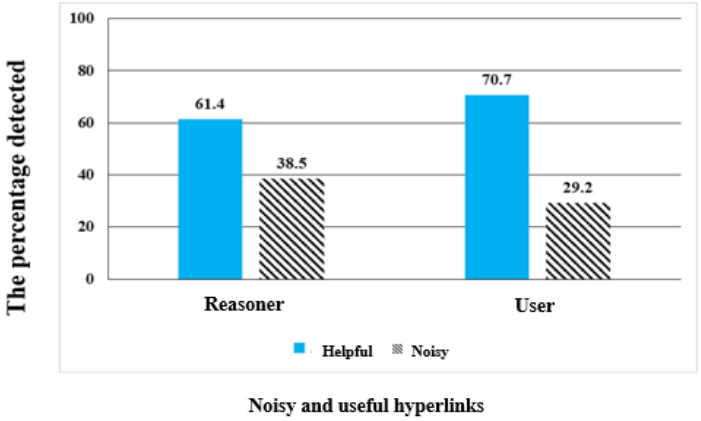


Figure 4. Percentage of hyperlink types as perceived by the user and the reasoned.

Conclusion

The proposed system takes the constructed dataset as input. Each row of the dataset consists of the class mapped from the hyperlink context topic of the source page, the class mapped from the topic of the target page, a field indicating the noisy or useful nature of the hyperlink from the user's perspective, and the domain name of the source page. The results were then compared with those of the user to show the extent to which each semantic or relational property in the ontology contributes to identifying a hyperlink as either noisy or useful. The categories of queries that lead users to noisy hyperlinks and the domains with the highest number of noisy hyperlinks were also identified. The experiments in the present study demonstrated the accuracy, capability and scalability of Semantic Web technologies in eliminating noisy hyperlinks.

کارافن
فصلنامه علمی دانشگاه فنی و حرفه‌ای
پاییز ۱۴۰۲، دوره ۲۰، شماره ۳، ۵۰۷-۴۸۵
آدرس نشریه: https://karafan.tvu.ac.ir/
doi:10.48301/KSSA.2023.382583.2426

پیاده‌سازی سیستم حذف ابرپیوندهای نویزی با استفاده از رویکرد معنایی و رابطه‌ای آنتولوژی DBpedia

کاظم تقندیکی[1*]

۱- عضو هیات علمی گروه مهندسی کامپیوتر، دانشگاه فنی و حرفه‌ای، تهران، ایران.

اطلاعات مقاله

نوع مقاله: مقاله پژوهشی

دریافت مقاله: ۱۴۰۱/۱۱/۰۷
بازنگری مقاله: ۱۴۰۲/۰۲/۱۶
پذیرش مقاله: ۱۴۰۲/۰۳/۲۱

کلید واژگان:
وب معنایی
ابرپیوندهای نویزی
آنتولوژی
استدلال‌گر
شباهت معنایی
شباهت رابطه‌ای

***نویسنده مسئول:** کاظم تقندیکی
پست الکترونیکی:
ktaghandiki@tvu.ac.ir

چکیده

همان‌طور که داده‌های وب به سرعت در حال گسترش و رشد هستند، ساختار گراف وب که یک نمایش گرافیکی از دنیای وب است، در حال بزرگ شدن می‌باشد و به تدریج ساختار محتوایی خود را به یک ساختار غیر محتوایی تبدیل کرده است. وجود داده‌های هرز مانند ابرپیوندهای نویزی در گراف ساختار وب، بسیاری از الگوریتم‌های لینک‌کاوی را با مشکل مواجه ساخته و باعث کاهش سرعت و بازدهی الگوریتم‌های بازیابی اطلاعات گردیده است. کارهای انجام شده به حذف ابرپیوندهای نویزی با استفاده رویکردهای ساختاری و رشته‌ای پرداخته‌اند. این رویکردها به اشتباه برخی از ابرپیوندهای مفید را حذف کرده و در بعضی شرایط قادر به تشخیص ابرپیوندهای نویزی نمی‌باشند. در این مقاله، ابتدا توسط یک خزنده تعاملی یک مجموعه داده از ابرپیوندهای نویزی و مفید با استفاده از خزش وب سایت‌ها ایجاد شد. سپس از طریق رویکردهای وب معنایی و امکاناتی نظیرآنتولوژی DBpedia به ساختار معنایی و رابطه‌ای این ابرپیوندها توجه گردید. در ادامه با فعال کردن استدلال‌گر آنتولوژی DBpedia، فرآیند حذف ابرپیوندهای نویزی از گراف ساختار وب صورت گرفت. آزمایش‌های انجام گرفته بر روی این سیستم، دقت و توانایی تکنولوژی‌های وب معنایی را در حذف ابرپیوندهای نویزی نشان می‌دهد.

شاپای الکترونیکی: ۲۵۳۸-۴۴۳۰
شاپای چاپی: ۲۳۸۲-۹۷۹۶

مقدمه

آزادی عمل در ایجاد صفحات وب و حجم زیادی از داده‌های ناساختاریافته در حوزه‌های مختلف تکنولوژی، باعث ایجاد یک چالش به نام «کلان داده‌ها» شده است. براساس آمار سال ۲۰۲۲ حدود ۲ میلیارد وب سایت توسط خزنده‌ها شاخص‌شده‌اند [۱]. انتشار صفحات وب در اینترنت در دهه‌های اخیر رشد چشم‌گیری داشته است. این افزایش داده در اینترنت باعث رشد گراف ساختار وب شده است. رشد بی‌رویه و وجود داده‌های هرز مانند ابرپیوندها در این گراف، پیمایش و کاوش مبتنی بر ساختار وب را دشوارتر کرده است [۲]. لذا باید این گراف را از ابرپیوندهای هرز پاک کرد. در ابتدا بسیاری از الگوریتم‌های بازیابی اطلاعات از محتویات اسناد وب برای دسته‌بندی، خوشه‌بندی و حذف صفحات هرز استفاده می‌کرده‌اند که حجم پردازشی و زمان زیادی مورد نیاز بود. شرکت‌ها به ارائه الگوریتم‌هایی در گراف ساختار وب پرداختند تا از ویژگی‌های ابرپیوندها به جای محتوایات اسناد استفاده کنند. به عنوان مثال فرایند دسته‌بندی اسناد وب، وابسته به ویژگی‌های ابرپیوند انجام گیرد و نه محتوایات اسناد. الگوریتم‌هایی مانند PageRank [۳] از همین ویژگی‌های ابرپیوندها در گراف ساختار وب استفاده کرده‌اند به طوری که حجم پردازش را در عملیات موتورهای جست‌وجو کاهش داده‌اند. اما پیش‌فرض اولیهٔ تمامی این الگوریتم‌ها این است که ابرپیوندها دقیقاً به همان صفحاتی اشاره می‌کنند که مد نظر کاربر است [۴]. لذا بسیاری از الگوریتم‌های لینک‌کاوی به اشتباه گراف ساختار وب را یک گراف کاملا محتوایی و معنایی در نظر می‌گیرند در حالی که این طور نیست. این گراف شامل یک سری ابرپیوندهای بی فایده و هرز می‌باشد که نه تنها کاربر را از هدفش دور می‌سازد، بلکه خروجی الگوریتم‌های بازیابی اطلاعات را با مشکل مواجه می‌کنند. در واقع وجود یک لینک یا ابرپیوند هرز می‌تواند این اجازه را به یک سند نامطلوب وب بدهد تا در رتبه‌ای بالاتر از سایر اسناد مناسب قرار بگیرد. این روش نامناسب افزایش رتبه، با مداخله انسان‌ها شکل می‌گیرد که مهمترین انگیزه شکل-گیری آن «کسب‌وکار» است [۴]. بعدها مطالعاتی برای کشف این نوع ابرپیوندها و حذف آنها از گراف ساختار وب انجام شد که بسیار وابسته به خصوصیات رشته‌ای [۴]، ساختاری [۵] ابرپیوندها می‌باشد و هیچ توجهی به ساختار معنایی و رابطه‌ای ابرپیوندها نمی‌شد. در این مقاله به ساختار معنایی و رابطه‌ای ابرپیوندها در دو سطح صفحه و سایت توجه شده و از تکنولوژی‌های وب معنایی مانند آنتولوژی[1]ها و استدلال‌گر[2]ها برای حذف ابرپیوندهای نویزی استفاده شده است. روش کار به این صورت است که ابتدا یک مجموعه داده از ابرپیوندها، در یک فرایند مجزا ساخته می‌شود، سپس با استفاده از تکنولوژی‌های وب معنایی مانند آنتولوژی‌ها و استدلال‌گرها، به آنالیز معنایی و رابطه‌ای مفهوم ابرپیوندها صفحه مبدأ و مفهوم صفحهٔ مقصد پرداخته خواهد شد، این آنالیز می‌تواند نویزی و مفید بودن ابرپیوندها را تشخیص دهد. نوآوری اصلی رویکرد پیشنهادی استفاده از ابزار مالت و دانش آنتولوژی DBpedia در شناسایی و حذف ابرپیوندهای نویزی می‌باشد که در بخش‌های بعدی این مقاله کامل توضیح داده شده‌اند.

سیستم طراحی شده یک ورودی از مجموعه داده ساخته شده، که هر سطر آن تشکیل شده است از کلاس نگاشت شده از موضوع حوزهٔ ابرپیوند صفحه مبدأ، کلاس نگاشت شده از موضوع صفحهٔ مقصد، نویزی یا مفید بودن ابرپیوند از دیدگاه کاربر و در نهایت نام دامنهٔ صفحه مبدأ را دریافت می‌کند، سپس بعد از تشخیص نویزی و مفید بودن ابرپیوندها، آن را با نتایج کاربر مقایسه می‌کند و مشخص خواهد کرد که هر کدام از ویژگی‌های معنایی و رابطه‌ای موجود در آنتولوژی به چه میزان در تعیین مفید و نویزی بودن ابرپیوندها نقش داشته‌اند. همچنین نشان خواهد داد، که کدام دسته از پرس‌وجوها، کاربر را به سوی ابرپیوندهای نویزی هدایت کرده است و این‌که کدام دامنه‌های وب، بیشترین ابرپیوندهای نویزی را داشته‌اند. و بسیاری از نتایج مفید دیگر. آزمایش‌های انجام گرفته بر روی این سیستم دقت و توانایی تکنولوژی‌های وب معنایی را در حذف ابرپیوندهای نویزی نشان می‌دهد.

[1] Ontology
[2] Reasoner

در ادامه، مقاله در چندین بخش سازماندهی شده است، بخش ۲، به بررسی کارهای قبلی در حذف ابرپیوندهای نویزی می‌پردازد. بخش ۳، به پیاده‌سازی رویکرد پیشنهاد شده خواهد پرداخت. بخش ۴، به بررسی آزمایش‌ها و نتایج به‌دست آمده می‌پردازد و در بخش ۵ به شرح یک نتیجه‌گیری کلی از رویکردمان خواهیم پرداخت.

کارهای انجام شده

کار [۶] ابرپیوندهای پیمایشی، تبلیغاتی و نامرتبط را ابرپیوند نویزی و دیگر ابرپیوندهای صفحه وب را ابرپیوند مفید در نظرگرفته‌اند. آنها از یک الگوریتم دسته‌بندی SVM که شامل دو کلاس شایسته و ناشایسته بود، برای شناسایی و فیلترکردن ابرپیوندهای نویزی استفاده کرده‌اند. در این کلاسیفایر برای پیداکردن ابرپیوندهای نویزی از شش معیار شباهت رشته‌ای استفاده شده بود. آنها کلاسیفایر خود را بر روی ابرپیوندهای ۱.۲ میلیون صفحه وب آزمایش کرده‌اند که در پایان شاهد حذف ۲۳٪ لینک‌ها بوده‌اند. در این سیستم از هیچ رویکرد معنایی و رابطه‌ای برای حذف ابرپیوندها استفاده نشده است. کار [۷] یک سیستم Anchor Woman را طراحی کرده‌اند که براساس ساختار ابرپیوندهای یک صفحه وب به تشخیص نویزی بودن ابرپیوندها می‌پرداخت. در سیستم طراحی شده ابرپیوندهای نویزی به ۳ دسته تقسیم می‌شوند که عبارتند از حلقه‌ای، چندتایی و بازگشت به خود. فرآیند کاری این سیستم به این شکل بود که ابتدا یک آدرس وب سایت را دریافت کرده و براساس یک جستجوی اول سطحی از ابرپیوندهای آن، ابرپیوندهای نویزی ذکر شده در بالا کشف و حذف می‌شدند. کار [۸] ابرپیوندهای نویزی را در سطح سایت تشخیص داده‌اند. آنها معتقد هستند که بین سایت‌ها ممکن است دو نوع رابطهٔ هرزی وجود داشته باشد، که عبارتند از: وجود رابطهٔ قویاً متصل بین فقط دو سایت، وجود رابطهٔ قویاً متصل بین زنجیره‌ای از سایت‌ها. آنها یک مقدار حد آستانه در نظر گرفته‌اند اگر لینک‌های تبادل شده بین سایت‌ها از این مقدار تجاوز کند، آن‌گاه تمامی ابرپیوندهای بین سایت‌ها نویز در نظر گرفته می‌شوند. الگوریتم ارائه شده توسط آنها در پایان ۱۶.۷٪ از لینک‌ها را با میانگین دقت ۵۹.۱۶٪ حذف کرد. آنها از یک رویکرد ساختاری برای حذف ابرپیوندهای نویزی استفاده کرده بودند. کار [۲] برای بهبود گراف ساختار وب و پیمایش راحت‌تر کاربران در وب سایت‌های مجموعهٔ دانشگاه‌های ایالت متحده آمریکا، از روش‌های مبتنی بر گراف استفاده کرده‌اند. آنها پس از استخراج بیش از ۶ میلیون لینک ارتباطی از ساختار وب ۱۱۰ دانشگاه در ایالات متحده، لینک‌ها را در پایگاه داده ذخیره کرده‌اند و با استفاده از ابزار Text Pipe بسیاری از لینک‌های نامطلوب موجود را در صفحات دانشگاه که به تصاویر، ویدیو، صدا مرتبط بوده‌اند حذف کرده‌اند. در پایان آنها شاهد بهینه شدن تعداد اسناد وب، طول مسیر و SCC در گراف ساختار وب بوده‌اند. کار [۹] با انجام فرآیند لینک‌کاوی بر روی گراف ساختار وب، یک تزاروس (مجموعه‌ای از کلمات معادل و مترادف) را با هدف افزایش دقت در سیستم پرس‌وجو موتورهای جست‌وجو ایجاد کرده‌اند. مراحل ساخت تزاروس از شش مرحله تشکیل شده است که مرحلهٔ دو آن به حذف ابرپیوندهای نویزی می‌پردازد. آنها در پایان شاهد حذف ابرپیوندهای پیمایشی با دقت ۹۲.۸۹٪ و افزایش دقت ۲۰٪ سیستم پرس‌وجو در موتورهای جست‌وجو بوده‌اند. رابطهٔ تجمعی در این پژوهش یک رویکرد معنایی می‌باشد، ولی رابطهٔ انجمنی را نمی‌توان یک رویکرد رابطه‌ای کامل در نظر گرفت. اگر چه در [۱۰] از رابطهٔ انجمنی یا افقی برای نشان دادن شباهت رابطه‌ای بین دو مفهوم استفاده کرده‌اند. ولی رابطهٔ افقی تنها نشان دهنده وجود یک رابطه Part OF بین دو مفهوم می‌باشد. درحالی که برای نشان دادن شباهت رابطه‌ای بین دو مفهوم از ویژگی‌های رابطه‌ای دیگری مانند Object Property ها در آنتولوژی‌ها، نیز می‌توان استفاده کرد که در این مقاله به آن توجه شده است. الگوریتم WSE ارائه شده توسط کار [۱۱] یکی دیگر از کارهای انجام شده در حذف ابرپیوندهای نویزی می‌باشد. مهمترین هدفی که ایشان دنبال می‌کرد، حفظ ابرپیوندهای معنایی و حذف ابرپیوندهای نویزی بود. ولی منظور ایشان از معنا تمرکز بر ساختار مسیری ابرپیوندها بود به گونه‌ای که ساختار سلسله مراتبی در ساختار ابرپیوندها حفظ شود و نه معنا یا هدف ابرپیوند. به عنوان مثال، چهار صفحه A، B، C و D را در نظر

بگیرید ابرپیوندهایی که A را به B و B را به C و C را به D وصل می‌کند، ابرپیوندهایی معنایی و ابرپیوندهایی که ارتباط بین صفحات را خلاف مسیر ذکر شده انجام می‌دهند، ابرپیوندهای نویزی به شمار می‌رود.

از جمله معایب کارهای انجام شده، تمرکز بر روی ابرپیوندهای سطح صفحه می‌باشد در حالی که امروزه بسیاری از سایت‌های اسپم، برای گرفتن رتبه بالا در عملیات شاخص‌گذاری گوگل از ابرپیوندهای سطح سایت استفاده می‌کنند و به دنبال ایجاد یک لینک نامشروع با دیگر سایت‌ها می‌باشد. در این مقاله، سعی شده است تا فرایند حذف ابرپیوندهای نویزی بر روی هر دو سطح صفحه و سایت نیز پرداخته شود. همچنین کارهای فوق تنها از رویکردهای رشته‌ای و ساختاری برای حذف ابرپیوندهای نویزی استفاده کرده‌اند. این رویکردها اگرچه دارای سرعت بالایی در عملیات تشخیص نوع ابرپیوند بودند اما به اشتباه برخی از ابرپیوندهای مفید را حذف کرده و در بعضی شرایط قادر به تشخیص ابرپیوندهای نویزی نمی‌باشند. به عنوان مثال در رویکرد رشته‌ای اگر ابرپیوندی دارای متن Bank (بانک) باشد که به صفحهٔ در مورد Bank (ساحل) اشاره می‌کند، یک ابرپیوند مفید در نظر گرفته می‌شود در حالی که رویکرد معنایی به کمک اطلاعات حوزهٔ متن ابرپیوند و صفحه مقصد به خوبی این ابرپیوند را نویز تشخیص می‌دهد و آن را حذف می‌کند. در این مقاله سعی شده است تا از طریق رویکردهای وب معنایی و امکاناتی نظیر آنتولوژی DBpedia به ساختار معنایی و رابطه‌ای ابرپیوندها توجه شود و با فعال کردن استدلال‌گر آنتولوژی DBpedia، فرایند حذف ابرپیوندهای نویزی انجام گیرد. آزمایش‌های انجام گرفته بر روی این سیستم دقت و توانایی تکنولوژی‌های وب معنایی را در حذف ابرپیوندهای نویزی نشان خواهد داد. جدول ۱ مقایسه‌ای بین نوع رویکرد ارائه شده و نوع ابرپیوندهای حذف شده کارهای انجام شده فوق را نشان می‌دهد.

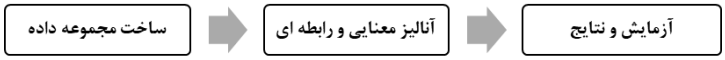
جدول ۱. مروری بر رویکرد و نوع ابرپیوندهای کارهای انجام شده.

کارهای انجام شده	الگوریتم شباهت	نوع ابرپیوندهای قابل تشخیص	دقت
کار [۶]	شباهت رشته ای	سطح صفحه	٪۷۸
کار [۷]	شباهت ساختاری	سطح صفحه	٪۸۲
کار [۸]	شباهت ساختاری	سطح سایت	٪۵۹.۱۶
کار [۲]	شباهت ساختاری	سطح صفحه	٪۷۸
کار [۹]	شباهت ساختاری و معنایی	سطح سایت	٪۹۲.۸۹
کار [۱۱]	شباهت ساختاری	سطح صفحه	٪۸۵.۳۷

کارهای فوق و کارهایی مانند [۲; ۱۱-۱۴] برای حذف ابرپیوندهای نویزی از مجموعه داده‌های موجود در بازیابی اطلاعات استفاده نکرده‌اند بلکه نویسندهٔ مقاله از طریق آنالیز گراف ساختاری یک وب سایت، به حذف ابرپیوندهای نویزی پرداخته است. لذا مراحل ساخت مجموعهٔ داده ابرپیوند مورد نیاز را پژوهشگر بر عهده گرفته است. در ادامه چگونگی ساخت این مجموعه داده با در نظر گرفتن معیارهای ساخت یک مجموعه داده در سیستم بازیابی اطلاعات [۱۵] شرح داده خواهد شد.

رویکرد پیشنهادی

پیاده‌سازی سیستم معنایی و رابطه‌ای حذف ابرپیوندهای نویزی دارای ۳ گام کلی می باشد که در شکل ۱ نشان داده شده است.



شکل ۱. فرآیند رویکرد پیشنهادی.

ساخت مجموعه داده

گام ساخت مجموعه داده، خود یک فرایند جدا می‌باشد که از تعدادی زیرگام تشکیل شده است. شکل ۲ مراحل ساخت مجموعه داده را نشان می‌دهد.

نوآوری اصلی رویکرد پیشنهادی در فازهای تشخیص موضوع و نگاشت موضوع به کلاس‌های آنتولوژی DBpedia می‌باشد. آنتولوژی DBpedia مجموعه‌ای از کلاس‌ها و خصوصیات است که محتوای ساختار یافته نمودار دانش وب سایت DBpedia را توصیف می‌کند. آنتولوژی DBpedia بر اساس استانداردهای وب معنایی پرکاربرد از جمله RDF و OWL است که اطلاعات ویکی‌پدیا را در مجموعه‌ای از کلاس‌ها مانند افراد، سازمان‌ها، رویدادها، آلبوم‌های موسیقی، فیلم‌ها و مفاهیمی مانند کتاب‌ها و رنگ‌ها نگاشت می‌کند. با تعریف اصطلاحات و روابط استاندارد برای توصیف مفاهیم رایج، آنتولوژیDBpedia چارچوبی سازگار و ساختار یافته را ارائه می‌کند به طوری که ماشین‌ها را قادر می‌سازد تا اطلاعات موجود در پایگاه دانش را درک و پردازش (استنتاج) کنند. این امر ساخت برنامه‌های کاربردی با استفاده از داده‌های DBpedia در پردازش زبان طبیعی، بازیابی اطلاعات و سایر زمینه‌ها را برای محققان و توسعه دهندگان آسان‌تر می‌کند [۱۶].

در بخش تشخیص موضوع: نویسنده مجموعهٔ اطلاعات متنی جمع‌آوری شده از ابرپیوند صفحهٔ مبدأ و مقصد را به یک یا چند کلمه (تحت عنوان موضوع) با استفاده از ابزار مالت نگاشت می‌دهد. در حالی که در سایر کارها بلافاصله بعد از استخراج اطلاعات متنی به بررسی شباهت رشته‌ای اطلاعات متنی جمع‌آوری شده از ابرپیوند صفحه مبدأ و مقصد پرداخته شده است، درادامه این بخش کامل شرح داده شده است.

در بخش یکپارچه سازی موضوعی: در این گام نویسنده از ترکیب کتابخانه WS4J و الگوریتم جست‌وجو معنایی استفاده می‌کند تا هر موضوع به دست آمده از مرحله قبل را به یک کلاس یا مفهوم در آنتولوژی DBpedia نگاشت دهد، درادامه این بخش کامل شرح داده شده است.

لذا در پایان به جای استفاده از چندین خط متن استخراج شده از ابرپیوند صفحهٔ مبدأ و صفحهٔ مقصد، تنها به بررسی وجود رابطه یا معنا در کلاس‌ها یا مفاهیم به‌دست آمده با استفاده از آنتولوژی DBpedia پرداخته می شود. مثلا فرض کنید ۲ کلاس، ماهی و آب به ترتیب از اطلاعات ابرپیوند صفحهٔ مبدأ و اطلاعات صفحهٔ مقصد به دست آمده است. اگر در آنتولوژی DBpedia رابطهٔ «زندگی کردن» یا هر نوع رابطهٔ دیگری بین دو کلاس ماهی و آب وجود داشته باشد، می‌توان تصمیم گرفت که این دو کلاس با یکدیگر شباهت رابطه‌ای دارند. یا فرض کنید ۲ کلاس، خودرو و لاستیک به ترتیب از اطلاعات ابرپیوند صفحهٔ مبدأ و اطلاعات صفحهٔ مقصد به دست آمده است. اگر در آنتولوژی DBpedia، رابطهٔ « جزئی از» یا Subclass Of و Has Superclass بین دو کلاس خودرو و لاستیک وجود داشته باشد، می‌توان تصمیم گرفت که این دو کلاس با یکدیگر شباهت معنایی دارند. اما اگر هیچ نوع رابطه‌ای بین دو کلاس به‌دست آمده با استفاده از آنتولوژی DBpedia شناسایی نشد، آن‌گاه ابرپیوند یک ابرپیوند نویزی می‌باشد.

شکل ۲. مراحل ساخت مجموعه داده.

انتخاب دامنه

سؤال اساسی این است که چه سایت‌هایی را باید مورد خزش قرار داد به طوری که دارای ابرپیوندهای نویزی و مفید باشند. وجود ابرپیوندهای نویزی به این دلیل که بتوان با ویژگی‌های معنایی و رابطه‌ای موجود درآنتولوژی آنها را حذف کرد و وجود ابرپیوندهای مفید برای این‌که با استفاده از همان ویژگی‌های معنایی و رابطه‌ای موجود در آنتولوژی‌ها بتوان مفید بودن آنها را اثبات کرد. سایت‌های مورد خزش باید سایت‌هایی باشند که محتوای آنها در مورد مفاهیم مورد علاقهٔ کاربران وب باشد. خوشبختانه سرویس ترند گوگل[1] به سوال ما جواب داده است، با بررسی‌های انجام شده مشخص شد کاربران بیش‌تر پرس‌وجوهای خود را در حوزهٔ مفاهیمی مانند بازیگران سینما، شخصیت‌های مشهور ورزشی و موسیقی، ویدیو، مدل، ورزش، اخبار، پول، خریدهای آنلاین و تکنولوژی‌ها جدید انجام می‌دهند. جدول ۲ پرس‌وجوهای مورد علایق کاربران اینترنتی با استفاده از سرویس ترند گول در سال ۲۰۲۱ را نشان می‌دهد.

جدول ۲. پرس‌وجوهای مورد علاقهٔ کاربران در موتور جست‌وجو گوگل [۱۷].

اخبار	افراد	ورزش	الکترونیک
Afghanistan	Jenifer Lawrence	Real Madrid CF	iPhone 13
AMC Stock	Kim Kardashian	Chelsea F.C	Galaxy Z Flip4
COVID Vaccine	Julie Gayet	Paris Saint-Germain F.C.	Nexus Summit
Dogecoin	Tracy Morgan	FC Barcelona	Motorola Moto G Power

سازمان‌ها یا اشخاص با به‌دست آوردن علایق کاربران اینترنتی می‌توانند دو رویکرد متفاوت را در طراحی صفات وب در نظر بگیرند.

۱- طراحی برای ایجاد سایت‌های بد: سایت‌هایی ایجاد می‌کنند که بر حوزه‌های مورد علاقهٔ کاربر به طور ضعیف کار کنند ولی در پشت صحنه از طریق ابرپیوندهای موجود در آن سایت‌ها یک نوع کسب و کار ثانویه پر سود مانند هدایت کاربر به سایت‌های فروشگاهی یا غیر اخلاقی را انجام می‌دهند. این نوع سایت‌ها دارای ابرپیوندهایی نویزی می‌باشند که هیچ توجهی به خواسته کاربران وب نمی‌کنند.

۲- طراحی برای ایجاد سایت‌های خوب: سایت‌هایی ایجاد می‌کنند که اطلاعاتی با کیفیت در حوزه‌های مورد علاقهٔ کاربر را به او نشان دهند. این نوع سایت‌ها دارای ابرپیوندهایی مفید می‌باشند که به خواستهٔ کاربران وب، توجه می‌کنند.

ایدهٔ اصلی این نوع روش انتخاب دامنه، خزش آن دسته دامنه‌هایی است که توسط موتورهای جست‌وجو در پاسخ به پرس‌وجوهای مورد علاقهٔ کاربران وب بازیابی می‌شوند. دامنه‌های بازیابی شده دارای دو حالت کلی می‌باشند. یا برای کسب منفعت غیرقانونی دارای ابرپیوندهای نویزی می‌باشند (سایت‌های بد) یا این‌که برای رساندن کاربر به هدفش و کسب منفعت قانونی دارای ابرپیوندهای مفید (سایت‌های خوب) می‌باشد. انتخاب سایت‌های خوب و بد مشخص کنندهٔ یک نوع دسته‌بندی باینری می‌باشد، لذا اگر مجموعهٔ دادهٔ آموزشی ساخته شده دارای رکوردهایی از هر دو نوع برچسب خوب و بد باشد، دقت نهایی مدل در تشخیص نویزی و مفید بودن یک ابرپیوند را افزایش می‌دهد.

خزش[2] دامنه

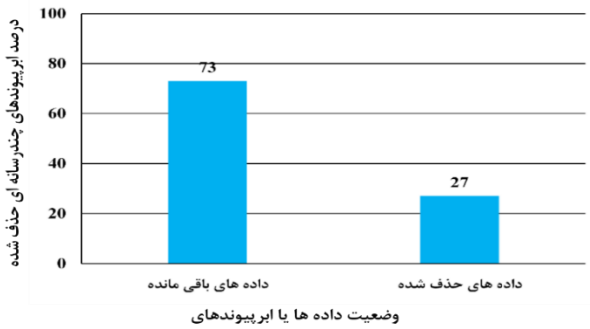
در این گام، کاربر هر کدام از مفاهیم به‌دست آمده از مرحلهٔ قبل را جداگانه در موتور جست‌وجو گوگل وارد کرده و از لیست وب سایت‌های بازیابی شده، تعدادی وب سایت را به طور تصادفی انتخاب می‌کند. در ادامه کاربر آدرس وب

[1] Google Trands
[2] Crawl

سایت‌ها را در یک سیستم خزندهٔ تعاملی که با کتابخانه‌های زبان برنامه‌نویسی جاوا نوشته شده است، وارد می‌کند. سپس خزنده شروع به خزش وب سایت کرده در نهایت لیستی از ابرپیوندهای آن دامنه را بازیابی می‌کند. لذا درکل با استفاده از سیستم خزنده ساخته شده صفحات مربوط به ۱۱۴ دامنه مفید و غیر مفید در باب موضوعات ذکر شده در جدول ۲، خزش شده‌اند. تمامی صفحات خزش شده دارای متن انگلیسی می‌باشند.

پیش‌پردازش[1] ابرپیوندها

از آنجا که رویکرد ارائه شده در گام پایانی از دانش استنتاج شده آنتولوژی DBpedia برای شناسایی و حذف ابرپیوندهای نویزی استفاده می‌کند، لذا تنها داده‌های رشته‌ای و متنی را به صورت ورودی دریافت می‌کند و قابلیت تشخیص نویزی یا هرزی بودن لینک‌های چندرسانه‌ای مانند تصاویر و ویدیو وجود ندارد. بنابراین ابرپیوندهای چندرسانه-ای در کنار ابرپیوندهای متنی تکراری نیز در این مرحله حذف شده‌اند. در ادامه با استفاده از دانش آنتولوژی DBpedia به شناسایی و حذف ابرپیوندهای نویزی متنی پرداخته شده است. کار [۲] تنها از همین روش برای شناسایی و حذف ابرپیوندهای نویزی نیز استفاده کرده است. در حالی که در رویکرد پیشنهادی از این روش تنها به عنوان یک پیش‌پردازش استفاده شده است و فرآیند اصلی حذف ابرپیوندهای نویزی با کمک دانش استخراج شده توسط استدلال‌گر آنتولوژی DBpedia انجام می‌گیرد. شکل ۳ وضعیت پیش‌پردازش ابرپیوندها را نشان می‌دهد.



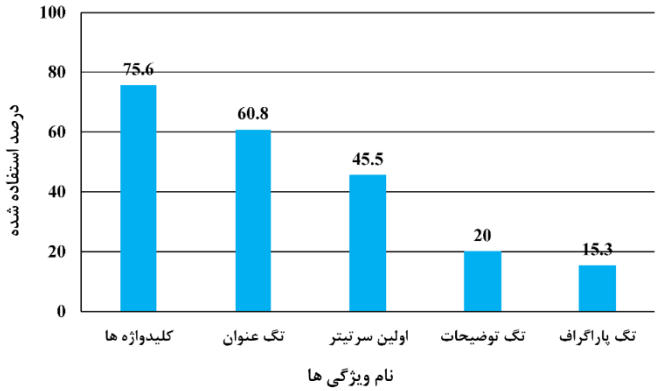
شکل ۳. پیش‌پردازش ابرپیوندهای چندرسانه‌ای.

شکل۳ نشان می‌دهد که ۲۷٪ از ابرپیوندها از نوع چند رسانه‌ای بوده که در گام پیش‌پردازش حذف شده‌اند.

استخراج ویژگی

برای این‌که سیستم در گام‌های بعدی بتواند با کیفیت خوبی عملیات تشخیص موضوع حوزهٔ متن ابرپیوند و صفحهٔ مقصد را انجام دهد، باید یک‌سری ویژگی‌های مناسب از صفحات وب استخراج شود تا از طریق این ویژگی‌ها، تشخیص موضوع انجام گیرد. اما باید بررسی کرد که کدام دسته از این ویژگی‌ها توسط طراحان وب، بیشترین کاربرد را در طراحی یک صفحهٔ وب دارند. شکل ۴ درصد کاربرد پنج تگ HTML کلیدی سازندهٔ صفحات وب را در ۵۰۰۰ صفحهٔ وب نشان می‌دهد؛ که سه ویژگی یا تگ کلیدواژه، عنوان صفحه و اولین سرتیتر بیشترین کاربرد را در طراحی صفحات وب دارند.

[1] Preprocessing



شکل ۴. وضعیت کاربرد ویژگی‌ها در طراحی صفحات وب.

لذا سه ویژگی عنوان صفحه، کلیدواژه‌ها و متن اولین سرتیتر صفحهٔ مقصد برای به‌دست آوردن موضوع صفحه مقصد استخراج شد و ویژگی‌های عنوان صفحه، کلیدواژه‌ها، متن ابرپیوند و پاراگرافی که ابرپیوند در آن قرار دارد توسط خزنده برای انتخاب موضوع حوزهٔ ابرپیوند از صفحه مبدأ استخراج گردید. جدول ۳و ۴ به ترتیب نمونه‌ای از ویژگی‌های استخراج شده برای تشخیص موضوع صفحهٔ مقصد و حوزهٔ متن ابرپیوند را نشان می‌دهند.

جدول ۳. ویژگی‌های استخراج شده برای تشخیص موضوع صفحهٔ مقصد از سایت فایل هیپو[1] [۱۸].

عنوان صفحه	Download free Software
کلید واژه‌ها	download software freeware shareware program
متن سرتیتر	Software

جدول ۴. ویژگی‌های استخراج شده برای تشخیص موضوع حوزهٔ متن ابرپیوند از سایت فایل هیپو [۱۸].

عنوان صفحه	Download free Software
کلید واژه‌ها	download software freeware shareware program
متن ابرپیوند	New Software
پاراگراف مربوط به ابرپیوند	The Latest Versions of the New Software

انتخاب ویژگی‌های فوق باعث افزایش سرعت عملیات آنالیز و تشخیص موضوع در گام‌های بعدی می‌شود در حالی که کارهای [۱۹-۲۱] از کل محتوای اسناد برای استخراج موضوع یک سند استفاده کرده‌اند.

پیش‌پردازش ویژگی

باید ویژگی‌های استخراج شده از مرحلهٔ قبل با کیفیت شوند، تا سیستم معنایی و رابطه‌ای با دقت بالا و خطای کمتری عملیات تشخیص موضوع را انجام دهد. در این مقاله برای افزایش کیفیت ویژگی‌های استخراج شده از عملیات

[1] FileHippo

پیش‌پردازش موجود در حوزهٔ متن‌کاوی مانند حذف stop words، token normalization و stemming [۱۵] استفاده شده است.

۱- **Stop words**: به آن دسته از کلمات، که به طور مکرر در نوشته‌ها تکرار می‌شوند Stop words گویند [۱۵].

۲- **Token Normalization**: فرآیندی استاندارد که به دنبال یکپارچه‌سازی واژه‌ها با شکل‌های متفاوت است [۱۵].

۳- **Stemming**: امروزه کاربران از فرم‌های متفاوت نوشتاری واژه‌ها در صفحات وب استفاده می‌کنند. فرآیند Stemming با حذف پسوندهای واژه‌ها، به دنبال تبدیل واژها به واژه ریشه‌شان می‌باشد.

برای انجام عملیات پیش پردازش در این مقاله، از کتابخانه spaCy زبان برنامه نویسی پایتون نیز استفاده شده است. شکل ۵ قطعه کد و خروجی فرآیند پیش پردازش را با استفاده از کتابخانه spaCy نشان می‌دهد.

قطعه کد

```
import spacy
nlp =
spacy.load("en_core_web_sm")
doc = nlp("Apple is looking at
buying U.K. startup for $1 billion")
for token in doc:
   print(token.text, token.lemma_,
   token.pos_, token.tag_,
   token.dep_, token.shape_,
   token.is_alpha, token.is_stop)
```

خروجی

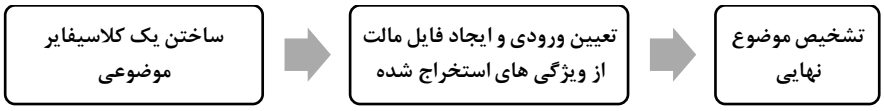
TEXT	LEMMA	POS	TAG	DEP	SHAPE	ALPHA	STOP
Apple	apple	PROPN	NNP	nsubj	Xxxxx	True	False
is	be	AUX	VBZ	aux	xx	True	True
looking	look	VERB	VBG	ROOT	xxxx	True	False
at	at	ADP	IN	prep	xx	True	True
buying	buy	VERB	VBG	pcomp	xxxx	True	False
U.K.	u.k.	PROPN	NNP	compound	X.X.	False	False
startup	startup	NOUN	NN	dobj	xxxx	True	False
for	for	ADP	IN	prep	xxx	True	True
$	$	SYM	$	quantmod	$	False	False
1	1	NUM	CD	compound	d	False	False
billion	billion	NUM	CD	pobj	xxxx	True	False

شکل ۵. فرآیند پیش پردازش با استفاده از ابزار spaCy.

ستون‌های LEMMA، TEXT، STOP در شکل ۵ به ترتیب فرآیندهای پیش‌پردازش ریشه‌یابی، جداسازی واژه‌ای و شناسایی ایست واژه‌ها را نشان می‌دهد.

تشخیص موضوع

زیرگام تشخیص موضوع به دنبال تعیین موضوع حوزهٔ متن ابرپیوند و صفحهٔ مقصد، توسط ویژگی‌های استخراج شده از فرآیند خزش (فاز قبلی) می‌باشد. تمامی عملیات مربوط به تشخیص موضوع توسط ابزار مالت انجام می‌گیرد. شکل ۶ فرایند یک تشخیص موضوع بانظارت را نشان می‌دهد. در ادامه هر کدام از گام‌های فرایند تشخیص موضوع شرح داده شده‌اند.

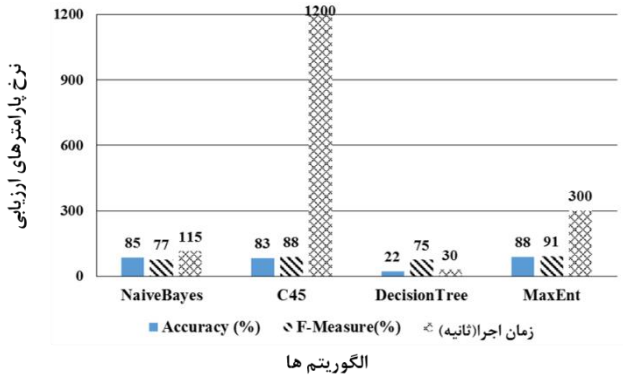


شکل ۶. فرایند تشخیص موضوعی با نظارت.

۱- ساخت یک کلاسیفایرموضوعی

برای ایجاد یک کلاسیفایر موضوعی باید یکسری ویژگی اولیه با کیفیت وجود داشته باشد، به همین دلیل تصمیم به ساخت یک دیتاست ویژگی گرفته شد. لذا ابتدا سه ویژگی کلیدی عنوان صفحه، کلیدواژه‌ها و سرتیتر صفحه که بیشترین کاربرد را در طراحی صفحات وب دارند، از آدرس ۵۰۰۰ صفحه وب استخراج شده و یک دیتاست ویژگی ساخته شد. برای ساخت یک کلاسیفایر، سیستم ۸۰٪ ازداده‌های دیتاست ویژگی را داده آموزشی در نظر می‌گیرد و از آنها برای ساختن یک کلاسیفایر موضوعی استفاده می‌کند. به همین ترتیب ۲۰٪ از داده‌ها را به عنوان دادهٔ آزمایشی در نظر می‌گیرد و از آنها برای ارزیابی دقت کلاسیفایر استفاده می‌کند.

در این مقاله برای ساخت یک کلاسیفایر موضوعی، از چهار روش دسته‌بندی Naïve Bayes، C45، Decision Tree و Max Entropy با Cross-Validation 10– [۲۲] جهت به‌دست آوردن یک موضوع خوب استفاده شد تا با توجه به دقت به‌دست آمده از هر کدام در هنگام تشخیص موضوع داده‌های آزمایشی، بهترین الگوریتم دسته‌بندی انتخاب شود. شکل ۷ نتایج حاصل از بررسی چهار روش دسته‌بندی فوق را نشان می‌دهد که با توجه به نتایج حاصل از آن، الگوریتم MaxEntropy دارای بهترین دقت تشخیص موضوع بر روی داده‌های آزمایشی می‌باشد.



شکل ۷. مقایسهٔ سه الگوریتم Naive Bayes، C45، Decision Tree و MaxEntropy.

۲- تعیین ورودی و ایجاد فایل مالت

در این گام، ابتدا توسط دستورهای خط فرمان، تمامی ویژگی‌های استخراج شده از حوزهٔ متن ابرپیوند و صفحهٔ مقصد به منظور تشخیص موضوع در اختیار ابزار مالت قرار گرفت. خروجی این دستورها، دو فایل خروجی با فرمت مالت می‌باشد که به آنها فایل بردار ویژگی نیز گفته می‌شود. فایل بردار ویژگی معمولاً یک نمایش عددی از ورودی‌ها را برای سریع‌تر شدن عملیات آنالیز فراهم می‌آورد [۲۲]. سپس این فایل به عنوان ورودی گام تشخیص موضوع نهایی استفاده می‌شود.

۳- تشخیص موضوع نهایی

این گام، خروجی مالت مرحلهٔ قبلی را دریافت کرده و تشخیص موضوع نهایی را با کلاسیفایر ساخته شده بر روی آن انجام می‌دهد. همچنین از خروجی ایجاد شده دوباره می‌توان به عنوان یک مدل برای استدلال موضوع داده‌های ورودی جدید درگام‌های بعدی نیز استفاده کرد. در پایان این گام، کلاسیفایر موضوعی ساخته شده، عملیات تشخیص

موضوع نهایی را برای حدود ۸۱.۸٪ از ویژگی‌های استخراج شده با موفقیت انجام داد. خروجی‌های به‌دست آمده از گام تشخیص موضوع نهایی، شامل ۴ فایل متنی مجزا می‌باشد.

یکپارچه‌سازی موضوعی

این گام، خروجی‌های عملیات تشخیص موضوع که در تعدادی فایل متنی قرار گرفته‌اند را با یکدیگر ترکیب و یکپارچه می‌کند. تا به یک ورودی مناسب برای گام نگاشت موضوع‌ها تبدیل شود. در پایان این گام، یک فایل csv به شکل ۸ ایجاد می‌شود.

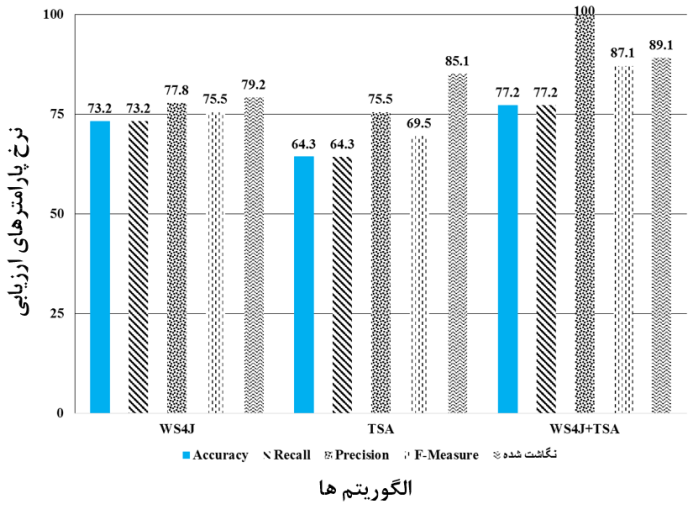
```
Bird,Fish,0,nationalgeographic
Bird,Mammal,0,nationalgeographic
Car,Canvas,1,dairyfoods
Song,Movie,1,songsmp3
```

شکل ۸. یکپارچه‌سازی موضوع ابرپیوندها و صفحات مقصد.

فیلد اول تا چهارم به ترتیب مشخص کنندهٔ موضوع حوزهٔ ابرپیوند صفحهٔ مبدأ، موضوع‌صفحهٔ مقصد، نویزی یا مفید بودن ابرپیوند از دیدگاه کاربر و در نهایت نام دامنهٔ صفحهٔ مبدأ می‌باشند. در واقع در این گام یک دیتاست موضوعی ساخته می‌شود.

نگاشت موضوع‌ها به کلاس‌ها

خروجی گام قبلی به عنوان ورودی گام نگاشت موضوع‌ها به کلاس‌های آنتولوژی DBpedia استفاده می‌شود. در این گام، هدف تبدیل موضوع ابرپیوندها و صفحات مقصد به کلاس‌هایی از آنتولوژی DBpedia می‌باشد. برای عملیات نگاشت از کتابخانه WS4J و الگوریتم جست‌وجو معنایی [۲۳] استفاده شد. برای بررسی دقت این دو الگوریتم از ۴۰۰ موضوع تصادفی استفاده شده است. شکل ۹ نتیجه پنج معیار ارزیابی دو الگوریتم فوق را بر روی ۴۰۰ موضوع نشان می‌دهد.



شکل ۹. مقایسه الگوریتم WS4J،جست‌وجو معنایی [۲۳] و ترکیب WS4Jوجست‌وجو معنایی.

با توجه به نتایج به‌دست آمده از دقت الگوریتم WS4J و الگوریتم جست‌وجو معنایی [۲۳]، مشخص شد که این دو الگوریتم می‌توانند مکمل یکدیگر باشند. لذا به بررسی دقت عملیات نگاشت ۴۰۰ موضوع با ترکیب این دو الگوریتم پرداخته شد. شکل ۹ همچنین نتیجه پنج معیار ارزیابی، حاصل از ترکیب این دو الگوریتم را در عملیات نگاشت نیز نشان می‌دهد. شکل ۱۰ نمونه‌ای از چند سطر تولید شده در آخرین گام فرایند ساخت مجموعه داده را نشان می‌دهد.

http://dbpedia.org/ontology/Automobile	http://dbpedia.org/ontology/Automobile	0	ebay
http://dbpedia.org/ontology/Currency	http://dbpedia.org/ontology/Actor	1	mydailyfundose
http://dbpedia.org/ontology/Continent	http://dbpedia.org/ontology/Actor	1	mydailyfundose
http://dbpedia.org/ontology/Activity	http://dbpedia.org/ontology/Actor	1	mydailyfundose

شکل ۱۰. نمونه‌ای از چندین سطر ایجاد شده در گام نگاشت.

فیلد اول تا چهارم به ترتیب مشخص کنندهٔ کلاس نگاشت شده از موضوع حوزهٔ ابرپیوند صفحهٔ مبدأ،کلاس نگاشت شده از موضوع صفحهٔ مقصد، نویزی یا مفید بودن ابرپیوند از دیدگاه کاربر و در نهایت نام دامنهٔ صفحهٔ مبدأ می‌باشد.

آنالیز معنایی و رابطه‌ای

مهم‌ترین مؤلفه در گام آنالیز معنایی و رابطه‌ای، استدلال‌گر می‌باشد. استدلال‌گر نرم‌افزاری است که بر روی یک یا تعدادی از دیتاست‌های مفهومی که با استفاده از آنتولوژی‌ها ساخته شده‌اند به کار گرفته می‌شوند تا نتایج منطقی را از حقایق موجود در آنتولوژی استخراج کند. در سیستم طراحی شده از استدلال‌گر Pellet استفاده شده است. استدلال‌گر Pellet با فعال شدن بر روی دانش اولیهٔ آنتولوژی DBpedia، یک دانش افزونه‌ای از ویژگی‌ها، رابطه‌ها وکلاس‌های آنتولوژی به‌دست می‌آورد. سپس یک سطر مجموعه‌داده را که نشان دهنده یک ابرپیوند از یک صفحهٔ وب می‌باشد به صورت دو ورودی مفهوم حوزهٔ متن ابرپیوند و صفحهٔ مقصد دریافت می‌کند و به بررسی وجود شباهت معنایی و رابطه‌ای بین آن دو مفهوم توسط ویژگی‌ها و رابطه‌های استدلال شده از آنتولوژی DBpedia می‌پردازد. از این پس مفهوم متن ابرپیوند را مفهوم مبدأ و مفهوم صفحهٔ مقصد را مفهوم مقصد نامیده می‌شود.

استدلال‌گر ابرپیوند ورودی را در صورتی مفید در نظر می‌گیرد که دارای حداقل یک ویژگی از سه ویژگی‌های زیر باشند:

۱- **Equivalent Class**: بین مفهوم مبدأ و مقصد ویژگی برابری وجود داشته باشد.

۲- **Subclass of و Has Superclass**: مفهوم مبدأ زیرکلاس یا سوپرکلاسی از مفهوم مقصد باشد.

دو رابطهٔ فوق را ویژگی‌های معنایی می‌گویند که نشان دهندهٔ شباهت معنایی بین دو مفهوم می‌باشد. به عنوان مثال مفهوم زن و فرد با یکدیگر شباهت معنایی دارند، چون زن یک زیر کلاسی از کلاس فرد است.

۳- **Object Property**: مفهوم مبدأ یا مقصد از طریق یک ویژگی شی با مفهوم دیگر رابطه داشته باشد. این نوع ویژگی را ویژگی رابطه‌ای می‌گویند که نشان دهنده شباهت رابطه‌ای بین دو مفهوم می‌باشد. به عنوان مثال دو مفهوم موز و میمون می‌توانند از طریق چند رابطه با یکدیگر شباهت رابطه‌ای داشته باشند، مانند دوست داشتن و خوردن (میمون، موز را دوست دارد یا میمون، موز را می‌خورد).

در رویکرد پیشنهادی نبودن حداقل یک ویژگی از ویژگی‌های مذکور، نشان‌دهندهٔ عدم شباهت معنایی و رابطه‌ای بین دو مفهوم مبدأ و مقصد می‌باشد. در این حالت سیستم معنایی، ورودی دریافتی را نویز تشخیص می‌دهد. لذا در پایان این گام دیتاستی مفهومی از ابرپیوندها ساخته می‌شود که نویزی و مفید بودن آن توسط استدلال‌گر مشخص شده است. شکل ۱۱ نمونه‌ای از سطرهای ساخته شده در گام آنالیز را نشان می‌دهد.

http://dbpedia.org/ontology/Automobile	http://dbpedia.org/ontology/Automobile	subClassOf	0	ebay
http://dbpedia.org/ontology/Cartoon	http://dbpedia.org/ontology/Agent	animator	0	wikipedia
http://dbpedia.org/ontology/Person	http://dbpedia.org/ontology/Place	birthPlace	0	wikipedia
http://dbpedia.org/ontology/Person	http://dbpedia.org/ontology/Place	livingPlace	0	wikipedia

شکل ۱۱. نمونه‌ای از چندین سطر تولید شده در گام آنالیز معنایی و رابطه‌ای.

فیلد اول تا پنجم به ترتیب مشخص کنندهٔ Subject، Object، ویژگی معنایی و رابطه‌ای تشخیص داده شده برای ارتباط فیلد اول و دوم توسط استدلال‌گر، نویزی و مفید بودن ابرپیوند از دیدگاه استدلال‌گر و در نهایت نام دامنه مربوط به صفحه مبدأ می‌باشند.

آزمایش و نتایج

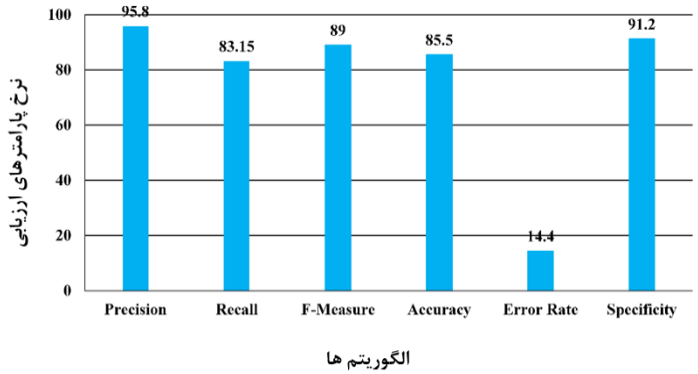
یکی از ابزارهای مفید که پژوهشگر را در ارزیابی رویکرد پیشنهادی کمک می‌کند، ماتریس آشفته نام دارد. ماتریس آشفته شامل دو روش برچسب‌دهی می‌باشد. یک نوع برچسب‌دهی، توسط سیستم در هنگام آنالیز معنایی و رابطه‌ای تعیین شده است و نوع دیگر برچسب‌دهی توسط کاربران خبره در هنگام ساخت مجموعه‌داده تعیین شده است. جدول ۵ ماتریس آشفته سیستم پیاده‌سازی شده را نشان می‌دهد. این جدول از آنالیزهای مربوط به کاربران در هنگام ساخت مجموعهٔ داده و سیستم حذف ابرپیوندهای نویزی با رویکرد معنایی و رابطه‌ای در هنگام آنالیز ساخته شده است. شکل ۱۲ نشان‌دهندهٔ مقدار آماری شش معیار ارزیابی پر کاربرد در سیستم بازیابی اطلاعات می‌باشد، که با استفاده از مقادیر جدول ۵ محاسبه شده‌اند. لذا رویکرد معنایی و رابطه‌ای پیشنهاد شده در برخورد با حذف ابرپیوندهای نویزی دارای دقت و قدرت تشخیص قابل قبول و نرخ خطای پایینی می‌باشد.

جدول ۵. ماتریس آشفته.

		برچسب‌های تشخیص داده شده توسط سیستم		
		YES	NO	Total
برچسب‌های واقعی تشخیص داده شده توسط کاربر	YES	(TP)۱۱۴۵	(FN) ۲۳۲	۱۳۷۷
	NO	(FP) ۵۰	(TN) ۵۱۹	۵۶۹
	Total	۱۱۹۵	۷۵۱	۱۹۴۶

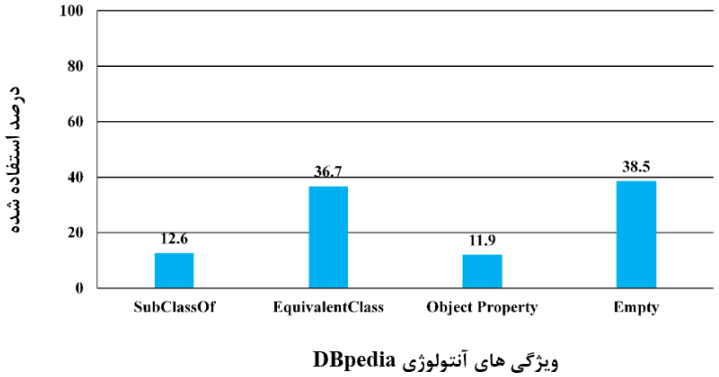
تعاریف انواع پارامترهای ارزیابی [۲۴]:

- **Accuracy**: نسبت نمونه‌های پیش‌بینی‌شده درست به تعداد کل نمونه‌های مجموعه داده را نشان می‌دهد.
- **Recall**: نسبت نمونه‌های مثبت (مثبت واقعی + منفی کاذب) را که به درستی توسط مدل شناسایی شده‌اند را نشان می‌دهد.
- **Precision**: نسبت پیش‌بینی‌های مثبت واقعی در تمام نمونه‌های پیش‌بینی شده مثبت را نشان می‌دهد.
- **F1-score**: میانگین هارمونیک دو معیار Precision و Recall را نشان می‌دهد.
- **Error rate**: نسبت پیش‌بینی‌های نادرست مدل را نسبت به تعداد کل نمونه‌ها نشان می‌دهد.
- **Specificity**: نسبت نمونه‌های منفی (منفی واقعی + مثبت کاذب) که به درستی توسط مدل شناسایی شده‌اند را نشان می‌دهد.



شکل ۱۲. نتایج شش معیار ارزیابی رویکرد معنایی و رابطه‌ای.

شکل ۱۳ نشان می‌دهد که استدلال‌گر به چه میزانی از انواع ویژگی‌های معنایی و رابطه‌ای موجود در آنتولوژی DBpedia برای نشان دادن شباهت معنایی و رابطه‌ای Subject و Object استفاده کرده است.



شکل ۱۳. وضعیت ویژگی‌های استفاده شده از انتولوژی DBpedia توسط استدلال‌گر.

شکل ۱۳ نشان می‌دهد که استدلال‌گر در ۱۲/۶٪ حالت، برای ارتباط‌سازی مفهوم ابرپیوند مبدأ به مفهوم صفحه مقصد از ویژگی SubclassOf استفاده کرده است. در ۳۶/۷٪ حالت، برای ارتباط‌سازی مفهوم ابرپیوند مبدأ به مفهوم صفحه مقصد از ویژگی EquivalentClass استفاده کرده است. پیدا کردن ویژگی‌های Subclass Of و Equivalent Class نشان‌دهندهٔ این است که بین مفهوم مبدأ و مقصد یک شباهت معنایی وجود دارد. همچنین استدلال‌گر در ۱۱/۹٪ حالت، برای ارتباط‌سازی مفهوم ابرپیوند مبدأ به مفهوم صفحهٔ مقصد از ویژگیObjectProperty نیز استفاده کرده است و در نهایت استدلال‌گر در ۳۸/۵٪ حالت، برای ارتباط‌سازی مفهوم ابرپیوند مبدأ به مفهوم صفحهٔ مقصد، هیچ ویژگی معنایی و رابطه‌ای مناسب به کمک آنتولوژی DBpedia پیدا نکرده است. جدول ۶ نمای کلی نتیجه استدلال‌گر را در آنالیز معنایی و رابطه‌ای را بر روی مجموعه داده نهایی(دیتاست مفهومی) نشان می‌دهد.

جدول ۶. نتایج کلی استدلال‌گر در آنالیز معنایی و رابطه‌ای بر روی مجموعهٔ داده

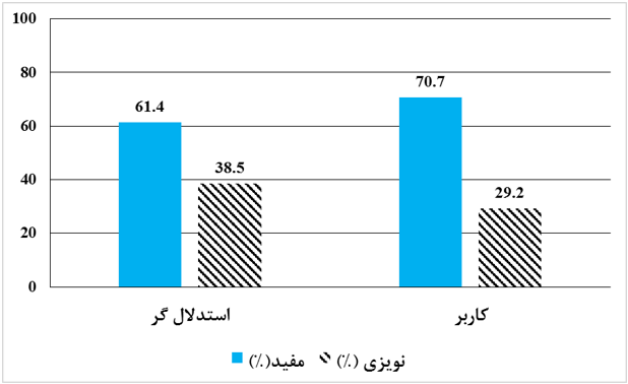
ابرپیوندهای خزش شده	ابرپیوندهای مجموعه داده
۲۶۶۵	۱۹۴۶
تعداد دامنه‌ها	**کلاس‌های انتولوژی**
۱۱۴	۳۱۲(۳۸٪)
دامنه‌های مفید	
Wikipedia, bbc, Facebook, YouTube, eBay	
دامنه‌های نویزی	
mydailyfundose, Google, fullmoviesfreedownload, hollywoodlife, wikipedia	
ابرپیوندهای مفید	**ابرپیوندهای نویزی**
۶۱٪/۴	۳۸٪/۵
مفاهیم مقصد یک ابرپیوند نویز	
Shopping Mall, Currency, Film, Actor, Drug	

برای درک بهتر نتایج آنالیز معنایی و رابطه‌ای نشان داده شده در جدول ۶، به مقایسهٔ آن با نتایج به‌دست آمده از دیدگاه کاربر نیز پرداخته شده است، جدول ۷ نمای کلی آنالیز مجموعه داده را از دیدگاه کاربر نشان می‌دهد که نویزی و مفید بودن ابرپیوندها را کاربر مشخص کرده است.

جدول ۷. آنالیز مجموعه داده از دیدگاه کاربر.

کل ابرپیوندهای خزش شده	ابرپیوندهای موجود در مجموعه داده
۲۶۶۵	۱۹۴۶
تعداد دامنه‌ها	
۱۱۴	
دامنه‌های مفید	
Wikipedia, bbc, Facebook, YouTube, eBay	
دامنه‌های نویزی	
mydailyfundose, Google, fullmoviesfreedownload, hollywoodlife, songmp3	
ابرپیوندهای مفید	**ابرپیوندهای نویزی**
۷۰٪/۷	۲۹٪/۲
مفاهیم مقصد یک ابرپیوند نویز	
Shopping Mall, Currency, Film, Actor, Model	

با مقایسهٔ جدول ۶ و ۷ به قدرت رویکرد معنایی و رابطه‌ای پیشنهاد شده، می‌توان پی برد. شکل ۱۴درصد ابرپیوندهای مفید و نویزی را از دو دیدگاه کاربر و استدلال‌گر نشان می‌دهد. از شکل ۱۴ می‌توان متوجه شد که رویکرد پیشنهادی همانند یک کاربر خبره، ابرپیوندهای نویزی و مفید را تشخیص می‌دهد.



نویزی و مفید بودن ابرپیوندها

شکل ۱۴. درصد ابرپیوندهای مفید و نویزی تشخیص داده شده از دیدگاه کاربر و استدلال‌گر.

دلایل بهینه بودن رویکرد پیشنهادی

جدول ۸ مقایسه بین کارهای انجام شده در بخش ۲ و رویکرد پیشنهادی را براساس ۳ معیار الگوریتم شباهت، نوع ابرپیوندهای قابل تشخیص و دقت را نشان می‌دهد.

۱- همان‌طور که در جدول ۸ نشان داده شده است رویکرد پیشنهادی از الگوریتم شباهت معنایی و رابطه‌ای دانش آنتولوژی DBpedia برای تشخیص و حذف ابرپیوندهای نویزی استفاده می‌کند، در حالی که کارهای انجام شده در بخش ۲، به رابطه و معنای موجود بین مفهوم ابرپیوند و صفحه مقصد توجهی نداشته و براساس شباهت رشته‌ای متن‌ها و ساختاری گراف‌ها به حذف ابرپیوندهای نویزی می‌پرداختند

۲- رویکرد پیشنهادی توانایی تشخیص ابرپیوندهای نویزی در دو سطح صفحه و سایت را دارد، در حالی که کارهای انجام شده در بخش ۲، تنها بر حذف تنها یکی از این نوع ابرپیوندهای نویزی تمرکز داشته‌اند.

۳- در رویکرد پیشنهادی مجموعه اطلاعات متنی استخراج شده برای ابرپیوند صفحهٔ مبدأ و مجموعه اطلاعات متنی استخراج شده برای صفحهٔ مقصد، تنها به یک موضوع (۱ یا ۲ کلمه) توسط ابزار مالت تبدیل می‌شدند، سپس با استفاده از الگوریتم‌های شباهت معنایی و WS4J به یک کلاس یا مفهوم از آنتولوژی DBpedia نگاشت پیدا می‌کردند. در نتیجه توسط دانش استخراج شده از آنتولوژی DBpedia توسط استدلال‌گر، وجود شباهت رابطه‌ای و معنایی بین مفاهیم بررسی می‌شد. در حالی که در کارهای انجام شده در بخش ۲، فرآیند بررسی شباهت رشته‌ای بر روی تمام اطلاعات متنی استخراج شده، انجام می‌گردید.

۴- رویکرد پیشنهادی دارای نرخ دقت ۸۵.۵۰٪ در دو سطح سایت و صفحه می‌باشد که در بین کارهای انجام شده دارای رتبه دوم است. در حالی که کار [۹] دقت ۹۲.۸۹٪ را تنها برای حذف ابرپیوندهای سطح سایت به‌دست آورده است.

جدول ۸. مقایسهٔ کارهای انجام شده در بخش ۲ و رویکرد پیشنهادی.

کارهای انجام شده	الگوریتم شباهت	نوع ابرپیوندهای قابل تشخیص	دقت
کار [۶]	شباهت رشته ای	سطح صفحه	٪۷۸
کار [۷]	شباهت ساختاری	سطح صفحه	٪۸۲

کارهای انجام شده	الگوریتم شباهت	نوع ابرپیوندهای قابل تشخیص	دقت
کار [۸]	شباهت ساختاری	سطح سایت	٪۵۹.۱۶
کار [۲]	شباهت ساختاری	سطح صفحه	٪۷۸
کار [۹]	شباهت ساختاری و معنایی	سطح سایت	٪۹۲.۸۹
کار [۱۱]	شباهت ساختاری	سطح صفحه	٪۸۵.۳۷
رویکرد پیشنهادی	شباهت معنایی و رابطه ای	سطح صفحه و سایت	٪۸۵.۵۰

نتیجه‌گیری و پیشنهادها

نوعی از داده‌های نویزی موجود در گراف ساختار وب، ابرپیوندها هستند. ابرپیوندهای نویزی، یک تأثیر منفی بر روی انواع الگوریتم‌های بازیابی اطلاعات می‌گذارند. بسیاری از این روش‌ها، بر روی ساختار رشته‌ای و گرافی ابرپیوندها متمرکز است و هیچ توجهی به ساختار معنایی و رابطه‌ای ابرپیوندها نکرده‌اند. لذا این رویکردها به اشتباه برخی از ابرپیوندهای مفید را حذف کرده و در بعضی شرایط قادر به تشخیص ابرپیوندهای نویزی نمی‌باشند. در این مقاله به ساختار معنایی و رابطه‌ای ابرپیوندها در دو سطح صفحه و سایت توجه شده و از تکنولوژی‌های وب معنایی مانند آنتولوژی‌ها و استدلال‌گرها برای حذف ابرپیوندهای نویزی استفاده شده است. روش کار به این صورت است که ابتدا یک مجموعه داده از ابرپیوندها، در یک فرایند مجزا ساخته می‌شود، سپس با استفاده از تکنولوژی‌های وب معنایی مانند آنتولوژی‌ها و استدلال‌گرها، به آنالیز معنایی و رابطه‌ای ابرپیوندها پرداخته خواهد شد. این آنالیز می‌تواند نویزی و مفید بودن ابرپیوندها را تشخیص دهد. سیستم طراحی شده یک ورودی از مجموعه دادهٔ ساخته شده را که هر سطر آن تشکیل شده است از کلاس نگاشت شده از موضوع حوزهٔ ابرپیوند صفحه مبدأ، کلاس نگاشت شده از موضوع صفحهٔ مقصد، نویزی یا مفید بودن ابرپیوند از دیدگاه کاربر و در نهایت نام دامنهٔ صفحهٔ مبدأ را دریافت می‌کند، سپس بعد از تشخیص نویزی و مفید بودن ابرپیوندها، آن را با نتایج کاربر مقایسه می‌کند و به خوبی نشان داد که هر کدام از ویژگی‌های معنایی و رابطهٔ موجود در آنتولوژی به چه میزان در تعیین مفید و نویزی بودن ابرپیوندها نقش داشته‌اند و این‌که کدام دسته از پرس‌وجوها، کاربر را به سوی ابرپیوندهای نویزی هدایت کرده است کدام دامنه‌های وب، بیشترین ابرپیوندهای نویزی را داشته‌اند و بسیاری از نتایجی دیگر. آزمایش‌های انجام گرفته بر روی این سیستم دقت و توانایی تکنولوژی‌های وب معنایی را در حذف ابرپیوندهای نویزی نشان داد .پژوهشگران می‌توانند سیستم پیاده‌سازی شده را در حوزه‌های پیشنهادی زیر توسعه و تعمیم دهند.

۱- آنتولوژی‌های متفاوتی را با آنتولوژی DBpedia استفاده شده در این سیستم، ترکیب می‌کنیم تا دامنهٔ بیشتری از مفاهیم سطح T-Box را پوشش دهد.

۲- از دیتاست‌های موجود در داده‌های فراپیوندی استفاده کرده تامفاهیم سطح A-Box را نیز پوشش دهد.

۳- عملیات تشخیص نویز با استفاده از آنتولوژی را از ابرپیوندهای متنی به تصاویر، ویدئو و صدا تعمیم خواهیم داد.

References

[1] Nalini, M. K., Dhinakaran, K., Elantamilan, D., Gnanavel, R., & Vinod, D. (2022, January 28-29). *Implementation of Indexing Techniques to Prevent Data Leakage and Duplication in Internet.* 2022 International Conference on Advances in Computing, Communication and Applied Informatics Chennai, India. https://doi.org/10.1109/ACCAI53970.2022.9752554

[2] Makkar, A., & Kumar, N. (2020). An efficient deep learning-based scheme for web spam detection in IoT environment. *Future Generation Computer Systems*, *108*, 467-487. https://doi.org/10.1016/j.future.2020.03.004

[3] Wu, Y., Wu, Y., Liu, Y., & Shi, T. (2022, March 25-27). *The research of the optimized solutions to Raft consensus algorithm based on a weighted PageRank algorithm.* 2022 Asia Conference on Algorithms, Computing and Machine Learning, Hangzhou, China. https://doi.org/10.1109/CACML55074.2022.00135

[4] Bhavitha, K. V., & Thangaraj, S. J. J. (2022, February 16-17). *Novel Detection of Accurate Spam Content using Logistic Regression Algorithm Compared with Gaussian Algorithm.* 2022 International Conference on Business Analytics for Technology and Security Dubai, United Arab Emirates. https://doi.org/10.1109/ICBATS54253.2022.9759003

[5] Benczur, A. A., Csalogany, K., Sarlos, T., & Uher, M. (2005, May 10-14). *Spamrank–fully automatic link spam detection work in progress.* Proceedings of the first international workshop on adversarial information retrieval on the web, Chiba, Japan. https://www.researchgate.net/publication/220846812_SpamRank_--_Fully_Automatic_Link_Spam_Detection

[6] Qi, X., Nie, L., & Davison, B. D. (2007, May 8). *Measuring similarity to detect qualified links.* Proceedings of the 3rd international workshop on Adversarial information retrieval on the Web, Banff, Alberta, Canada. https://doi.org/10.1145/1244408.1244418

[7] Wookey, L., & Geller, J. (2004). Semantic hierarchical abstraction of web site structures for web searchers. *Journal of Research and Practice in Information Technology*, *36*(1), 23-34. https://doi.org/10.3316/ielapa.120100890765820

[8] Da Costa Carvalho, A. L., Chirita, P. A., De Moura, E. S., Calado, P., & Nejdl, W. (2006, May 23-26). *Site level noise removal for search engines.* Proceedings of the 15th international conference on World Wide Web, Edinburgh, Scotland. https://doi.org/10.1145/1135777.1135793

[9] Chen, Z., Liu, S., Wenyin, L., Pu, G., & Ma, W-Y. (2003, August 1). *Building a web thesaurus from web link structure.* Proceedings of the 26th annual international Association for Computing Machinery SIGIR conference on Research and development in informaion retrieval, Toronto, Canada. https://doi.org/10.1145/860435.860447

[10] Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004, May 2-7). *WordNet::Similarity: measuring the relatedness of concepts.* Demonstration Papers at Human Language Technology-NAACL 2004, Boston, Massachusetts. https://doi.org/10.5555/1614025.1614037

[11] Li, F. (2008, October 12-14). *Extracting Structure of Web Site Based on Hyperlink Analysis.* 2008 4th International Conference on Wireless Communications, Networking and Mobile Computing, Dalian, China. https://doi.org/10.1109/WiCom.2008.2538

[12] Keller, M., & Nussbaumer, M. (2011, September 7-9). *Beyond the Web Graph: Mining the Information Architecture of the WWW with Navigation Structure Graphs.* 2011 International Conference on Emerging Intelligent Data and Web Technologies, Tirana, Albania. https://doi.org/10.1109/EIDWT.2011.23

[13] Zheng, Y., Cheng, X-C., & Chen, K. (2008). Filtering noise in Web pages based on parsing tree. *The Journal of China Universities of Posts and Telecommunications*, *15*(25), 46-50. https://doi.org/10.1016/S1005-8885(08)60153-3

[14] Bechhofer, S., Harmelen, F. V., Hendler, J., Horrocks, I., McGuinness, D. L., Patel-Schneider, P. F., & Stein, L. A. (2004). *OWL Web Ontology Language Reference*. W3C. https://www.w3.org/TR/owl-ref/

[15] Widyassari, A. P., Noersasongko, E., Syukur, A., & Affandy. (2022, December 8-9). *The 7-Phases Preprocessing Based On Extractive Text Summarization.* 2022 Seventh

International Conference on Informatics and Computing, Denpasar, Bali, Indonesia. https://doi.org/10.1109/ICIC56845.2022.10006998

[16] Rasham, S., Naz, A., Afzal, Z., Ahmed, W., Abbas, Q., Anwar, M. H., Ejaz, M., & Ilyas, M. (2022). The Challenges and Case for Urdu DBpedia. In A. Ullah, S. Anwar, Á. Rocha, & S. Gill (Eds.), *Proceedings of International Conference on Information Technology and Applications*. Springer Nature Singapore. https://doi.org/10.1007/978-981-16-7618-5_38

[17] GoogleTrends. (2021). *Explore what the worldthe world is searching for right now*. https://trends.google.com/trends/

[18] FileHippo. (2021). *FileHippo.com - Download Free Software*. https://filehippo.com/

[19] Ercan, G., & Cicekli, I. (2007). Using lexical chains for keyword extraction. *Information Processing & Management*, *43*(6), 1705-1714. https://doi.org/10.1016/j.ipm.2007.01.015

[20] Joshi, C., Attar, V. Z., & Kalamkar, S. P. (2022). An Unsupervised Topic Modeling Approach for Adverse Drug Reaction Extraction and Identification from Natural Language Text. In S. Tiwari, M. C. Trivedi, M. L. Kolhe, K. K. Mishra, & B. K. Singh (Eds.), *Advances in Data and Information Sciences*. Springer Singapore. https://doi.org/10.1007/978-981-16-5689-7_44

[21] Lott, B. (2012). *Survey of keyword extraction techniques*. UNM Education. https://www.docdroid.net/bii3/lott-pdf#page=10

[22] Fedorov, A. M., & Datyev, I. O. (2022). The Effect of Additive Regularization for Topic Modeling of Social Media Communities. In R. Silhavy (Ed.), *Artificial Intelligence Trends in Systems*. Springer International Publishing. https://doi.org/10.1007/978-3-031-09076-9_51

[23] Zaeri, A., & Nematbakhsh, M. A. (2012). A Terminological Search Algorithm for Ontology Matching. *Modern Applied Science*, *6*(10), 37-52. https://doi.org/10.5539/mas.v6n10p37

[24] Ahuja, R., Chug, A., Kohli, S., Gupta, S., & Ahuja, P. (2019). The Impact of Features Extraction on the Sentiment Analysis. *Procedia Computer Science*, *152*, 341-348. https://doi.org/10.1016/j.procs.2019.05.008