The Role of Different Types of Homograph Contexts in Measuring Documents Similarities

Hajar Sotudeh
PhD in Knowledge and Information Science;
Associate Professor; Shiraz University;
Corresponding Author sotudeh@shirazu.ac.ir

Mojgan Houshyar
MA in Knowledge and Information Sciences; Shiraz University;
mojganhoushyar1365@gmail.com

Iranian Journal of Information Processing and Management

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 33 | No. 3 | pp. 1183-1206
Spring 2018

Received: 06, Aug. 2016 **Accepted: 14, May 2017**

Abstract: Automatic information retrieval is based on the assumption that texts contain content or structural elements that can be used in word sense disambiguation and thereby improving the effectiveness of the results retrieved. Homographs are among the words requiring sense disambiguation. Depending on their roles and positions in texts, homograph contexts could be divided to different types, with probably different potency in determination of similarity of documents. Using a content analysis method, the present research aims to compare the powers of five kinds of contexts including text citations, references, reference titles, paper titles and texts in homograph sense disambiguation.

Applying a content analysis method, the present paper concentrates on a document test collection built on English homographs by choosing a sample consisted of 3637 articles containing 19 homographs about 54 subjects published during 2000-2015. Discriminant analysis was used to determine the similarity within or differentiation between the 54 document clusters.

According to the results of the discriminant analyses carried out within each of the clusters, sub-clusters of documents can be observed, though with a very little differentiation in terms of the homograph contexts. Text-citation and reference contexts are revealed to have minimum role in differentiating between the documents within the clusters.

Documents containing synonymous homographs form clusters within which documents are rather similar in terms of their homograph contexts. Furthermore, homograph context types are not equal in their power to determine similarities. Text-citation context and reference context types showed the highest degree of similarities within the clusters. These two context types, which show high similarity within clusters, can be used to improve retrieval results. It is suggested that the results of the comparison

of these two contexts can be used as a tool for secondary ranking or clustering of information retrieval results.
This is the first research of its kind to define different text contexts and compare them in terms of their power to determine similarity of texts containing synonymous homographs.

Keywords: Homographs, Similarity, Word Sense Disambiguation, Title Context, Reference-Title Context, Text-Citation Context, Text Context, Reference Context, Discriminant Analysis

بررسی نقش انواع بافتار هم‌نویسه‌ها در تعیین شباهت بین مدارک

هاجر ستوده
دکتری علم اطلاعات و دانش‌شناسی؛ دانشیار؛ دانشگاه شیراز؛
پدیدآور رابط sotudeh@shirazu.ac.ir

مژگان هوشیار
کارشناس ارشد علم اطلاعات و دانش‌شناسی؛ دانشگاه شیراز mojganhoushyar1365@gmail.com

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۳۹۵/۰۵/۱۶ **پذیرش:** ۱۳۹۶/۰۲/۲۴ مقاله برای اصلاح به مدت ۳ روز نزد پدیدآوران بوده است.

فصلنامه | علمی پژوهشی
پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISTA، ISC، و SCOPUS
jipm.irandoc.ac.ir
دورهٔ ۳۳ | شمارهٔ ۳ | ص‌ص ۱۱۸۳-۱۲۰۶
بهار ۱۳۹۷

چکیده: رفع ابهام معنایی واژگان در بازیابی خودکار اطلاعات از چالش‌های بزرگ این حوزه است. متن در خود عناصری محتوایی یا ساختاری دارد که با شناسایی و تحلیل آن‌ها و استخراج الگوهای متفاوت می‌توان به رفع ابهام معنایی و در نتیجه، افزایش اثربخشی نتایج بازیابی دست یافت. هم‌نویسه‌ها از جمله واژگانی هستند که به رفع ابهام معنایی نیاز دارند. نشان داده شده است که بافتار هم‌نویسه می‌تواند به بهبود بازیابی آن کمک کند. بافتار هم‌نویسه خود می‌تواند بسته به نقش و جایگاه آن در متن به انواعی تقسیم شود که ممکن است هر یک در تعیین شباهت بین مدارک از قدرت متفاوتی برخوردار باشد. هدف اصلی از این پژوهش، مقایسهٔ پنج نوع بافتار متنی (شامل بافتار استناد متنی، ارجاع، عنوان ارجاع، عنوان و متن مقاله) به لحاظ قدرت آن‌ها در تعیین شباهت میان مدارک است.

به کمک روش تحلیل متن، مجموعه‌ای آزمایشی از مدارک پیرامون هم‌نویسه‌های انگلیسی مشتمل بر ۳۶۳۷ مقالهٔ منتشرشده در بازهٔ زمانی ۲۰۰۰-۲۰۱۵ پیرامون ۱۹ هم‌نویسه در ۵۴ گروه موضوعی مورد بررسی قرار گرفت. برای تعیین شباهت درون خوشه‌ها از روش آماری تحلیل تشخیصی استفاده شده است.

نتایج تحلیل تشخیصی نشان داد که در درون خوشه‌های معنایی، زیرخوشه‌هایی با تمایز بسیار اندک قابل مشاهده است. دو بافتار استناد متنی و ارجاع کمترین نقش را در ایجاد تمایز و در نتیجه، بیشترین شباهت را در درون خوشه‌ها داشته‌اند.

نتایج به‌دست‌آمده نشان داد که هم‌معنا بودن هم‌نویسه‌ها به ایجاد خوشه‌هایی از مدارک منجر می‌شود که در درون آن‌ها مدارک با هم به لحاظ انواع بافتار هم‌نویسه‌ای تفاوت چندانی با هم ندارند. همچنین،

انواع بافتار از قدرت برابری در تعیین تشابه بین مدارک برخوردار نیستند. دو بافتار استناد متنی و ارجاع در تعیین شباهت معنایی در درون خوشه‌های معنایی بیشترین قوت را داشته‌اند. از این دو نوع بافتار که قوت بیشتری در ایجاد شباهت داشته‌اند، می‌توان برای بهبود نتایج بازیابی کمک گرفت. الگوریتم‌های بازیابی در موتورهای جست‌وجو و پایگاه‌های اطلاعاتی می‌توانند برای سنجش دقیق‌تر شباهت بین مدارک حاوی هم‌نویسه‌ها از تحلیل این دو نوع بافتار استفاده کنند.
اثر حاضر نخستین پژوهشی است که به تعریف انواع گوناگونی از بافتارهای متنی و مقایسهٔ آن‌ها به‌منظور تعیین قدرت آن‌ها در سنجش شباهت مدارک حاوی هم‌نویسه‌های هم‌معنا می‌پردازد.

کلیدواژه‌ها: هم‌نویسه، شباهت معنایی، بافتار عنوان، بافتار عنوان ارجاع، بافتار استنادی متنی، بافتار متن، تحلیل تشخیصی

۱. مقدمه

بازیابی اطلاعات ارتباط تنگاتنگی با مبحث معناشناسی دارد (لنکستر ۲۰۰۴؛ اینگورسن ۱۹۹۲؛ مهراد و فلاحتی فومنی ۱۳۸۴). از یک‌سو، کاربران در غالب اوقات به‌دنبال جست‌وجوی موضوع یا موضوع‌هایی خاص در سامانهٔ بازیابی اطلاعات هستند (زمانی ۱۳۸۷) و آشکار است که درک موضوع و محتوای اثر در گرو درک معنای آن از سوی کاربر است. از سوی دیگر، نمایه‌سازان پس از تعامل با متن به تفسیر آن می‌پردازند و سپس، اصطلاحاتی را به اثر تخصیص می‌دهند. تخصیص این اصطلاحات بر پایهٔ ادراکی صورت می‌گیرد که نمایه‌ساز از محتوای اثر، محتوای رده، و شباهت آن دو به‌دست می‌آورد (لنکستر ۲۰۰۴). از این رو، گفته می‌شود که بازیابی اطلاعات به شدت در سیطرهٔ معناست (پائو ۱۹۸۹؛ Fillmore 1985).

بازیابی اطلاعات در گرو تحقق مفهوم ربط است (ساراسویک ۲۰۱۱)، سامانه‌های خودکار بازیابی اطلاعات به‌دلیل تکیه بر نحو و ویژگی‌های لغوی از درک معنا و روابط بین واژگان عاجز هستند (اینگورسن ۱۹۹۲). از آنجا که شباهت نحوی و لغوی لزوماً شباهت معنایی را در پی ندارد (چامسکی ۲۰۰۰). تکیهٔ صرف بر شباهت نویسه‌ای باعث بازیابی حجم زیادی از مدارک نامرتبط یا کم‌ربط با نیاز اطلاعاتی کاربر می‌شود. در نتیجه، بازیابی اطلاعات اثربخشی لازم را نخواهد داشت. این امر به‌ویژه برای واژگان چندمعنایی (از

جمله هم‌نویسه‌ها[1]) دوچندان می‌شود. هم‌نویسه‌ها به واژگانی اطلاق می‌شود که شکل نوشتاری و تلفظ یکسانی دارند، ولی دارای معانی متفاوتی هستند (فلاحتی فومنی ۱۳۸۵؛ رخشانفر ۱۳۷۱؛ علوی‌مقدم ۱۳۸۹؛ بی‌جن‌خان و مرادزاده ۱۳۸۹ ;Harford & Heasly1983 ;Pallmer 1976). مانند "Bat" در دو معنای (خفاش و چوب بیسبال). واژگان هم‌نویسه یکی از مهم‌ترین لایه‌های ایجاد ابهام در زبان به شمار می‌آید (مسعودی، راحتی قوچانی و استاجی ۱۳۸۹الف) که بازیابی اطلاعات را با چالش‌های معنایی مواجه می‌سازد. به هر روی، هم‌نویسه‌ها در هر زبانی به‌دلیل معناهای متفاوتی که از آن‌ها استنباط می‌شود، موجب ابهام زیادی در درک متن و ریزش کاذب[2] شدید، به‌خصوص در جست‌وجوهای تخصصی می‌گردند (مینایی بیدگلی، اکبری و حسنی ۱۳۸۶). از آنجا که ریزش کاذب به بازیافت بالا و دقت کم می‌انجامد (یوسفی ۱۳۷۶)، برای کمک به کاربر در یافتن مدارک مرتبط، به الگوریتم‌ها و فنون خودکاری نیازمندیم که بتواند شاخص‌هایی را از خود متن استخراج کند و نیازی به تفسیر معنایی انسان و دانش پایه مانند اصطلاحنامه نداشته باشد (بی‌جن‌خان و مرادزاده ۱۳۸۹). خودکارسازی کامل فرایند بازیابی اطلاعات به‌احتمال قوی اثربخشی بازیابی را بهبود می‌بخشد. با وجود مزایایی که روش‌های خودکار در بازیابی اطلاعات دارند، اما درک معنا از چالش‌های اساسی فراروی این روش‌هاست. رایانه هنوز به مرحلهٔ ادراک «معرفتی» متن نرسیده است. بر خلاف انسان که با تحلیل بافتار پیرامون یک کلمه و به‌کمک دانش «همیشه حاضر»[3] خود به مفهوم واژه می‌رسد، رایانه از این دانش «همیشه حاضر» برخوردار نیست (Searle 1984; Weizenbaum 1976; Machlup 1983؛ اینگورسن ۱۹۹۲) و برای ادراک بافتار متن به ساختار دانشی که ساختهٔ بشر باشد، نیاز دارد.

باور بر آن است که متن، عناصر لازم و کافی برای کمک به بازیابی آن را در خود دارد (Hjørland 2008). از این رو، تلاش محققان بر آن است که به کمک ویژگی‌ها و نشانه‌های موجود در خود متن به درک معنای آن و در نتیجه، بازیابی اثربخش آن بپردازند (مسعودی، راحتی قوچانی و استاجی ۱۳۸۹ الف). عناصر متنی که تاکنون در پژوهش‌ها به‌عنوان شاخص‌هایی جهت درک معنا و در نتیجه، تمایز یا سنجش شباهت بین مدارک به‌کار رفته، عبارت است از بسامد واژه‌ها، هم‌رویدادی واژه‌ها، ادات سخن، وندها و جز

1. homographs 2. false drop 3. ubiqutious

آن. همگی این شاخص‌ها، عناصری از بافتار متن را تشکیل می‌دهند که عمدتاً از نوع ریخت‌شناختی، زبان‌شناختی یا دستوری بوده‌اند. آشکار است که هرچه ویژگی‌ها و شاخص‌های متنی بیشتری کشف شود و مورد آزمایش قرار گیرد، دسترسی به شاخص یا دسته شاخص‌هایی که از توان بیشتری در تمایز مدارک و آشکارسازی معنا برخوردار باشند، امکان‌پذیرتر خواهد شد. در این میان، اخص‌سازی نوع بافتار واژگانی می‌تواند بخش اعظمی از بار سامانه را بکاهد و کار ابهام‌زدایی معنایی را برای الگوریتم‌ها آسان‌تر سازد. نشان داده شده است که بافتار هم‌نویسه می‌تواند به بهبود بازیابی آن کمک کند. بافتار هم‌نویسه خود می‌تواند بسته به نقش و جایگاه هم‌نویسه در متن به انواعی تقسیم شود (بافتارهای هم‌نویسه‌ای عنوان[1]، ارجاع[2]، استناد متنی[3]، عنوان ارجاع[4] و متن مدرک[5]) که ممکن است هر یک در تعیین شباهت بین مدارک از قدرت متفاوتی برخوردار باشند.

پژوهش‌های پیشین اثربخشی بافتار هم‌نویسه‌ها در رفع ابهام معنایی از این نوع کلمات را تأیید کرده‌اند (مسعودی، راحتی قوچانی و استاجی ۱۳۸۹ الف؛ مسعودی، راحتی قوچانی و استاجی ۱۳۸۹ب؛ بزم‌آرا، معروفی و پیله‌ور ۱۳۹۰؛ بزم‌آرا، جعفری و بزم‌آرا ۱۳۹۲؛ امیدزاده، موسوی و نادری ۱۳۹۲؛ Hindl 1990 Hearst 1991; Lee Ng & Chia 2004; Tesprasit, Charoenpornsawat & Sornlertlamvanich 2003;). همچنین، نقش برخی از انواع بافتارها برای مثال، بافتار عنوان (Tang, Wang, & Zhang 2012 Han et al. 2004;)، بافتار استناد (Tsai, Kundu & Roth 2013; ; Nanba & Okumura1999; Schwartz & Hearst 2004; Nakov, Gazvinian & Radev 2010; Ding 2014; & Jeong, Song 2011; Small 2014; Xu & Lin Jbara Ezra & Radev 2013; Liu, Chen, Ding, Wang) و عنوان ارجاع (Choi 2010 2012; Tang et al.) در سنجش شباهت میان مدارک تأیید شده است. با این حال، نقش انواع بافتار به‌طور ویژه در رفع ابهام معنایی از هم‌نویسه‌ها مورد بررسی قرار نگرفته است. با آن‌که چندین پژوهش در حوزهٔ نقش بافتارها به عمل آمده، همچنان بر ضرورت انجام تحقیق در این باره تأکید می‌شود (کریمپور و همکاران ۲۰۰۹؛ مسعودی، راحتی قوچانی و استاجی ۱۳۸۹ الف). شاخص‌هایی مانند روابط بین مدارک از جمله استناد متنی، عنوان ارجاع، ارجاع، عنوان و متن مدارک که از انواع خاص بافتار هم‌نویسه به شمار می‌آیند، تاکنون در پژوهش‌های انجام‌شده

1. paper title context
2. body of reference context
3. text citation context
4. reference title context
5. text context

پیرامون رفع ابهام معنایی هم‌نویسه‌ها مورد توجه قرار نگرفته است. همچنین، پژوهشی که از این انواع بافتارها در رفع ابهام معنایی هم‌نویسه‌ها و همچنین، مقایسهٔ توان این بافتارها در تعیین شباهت مدارک حاوی هم‌نویسه‌های هم‌معنا بپردازد، صورت نگرفته است. بنابراین، پژوهش حاضر می‌کوشد با تمرکز بر هم‌نویسه‌های شناسایی‌شده در تحقیق Twilley et al. (1994) و ساخت مجموعه‌ای آزمایشی[1] متشکل از مجموعه مقالات منتشرشده پیرامون این هم‌نویسه‌ها در پایگاه «گوگل‌اسکالر» در بازهٔ زمانی ۲۰۰۰ -۲۰۱۵، نقش انواع بافتار هم‌نویسه‌ها را در تعیین شباهت بین این مدارک مورد بررسی قرار دهد. بدین منظور، مدارک حاوی هم‌نویسه‌های هم‌معنا در درون خوشه‌هایی قرار داده می‌شود و سپس، به کمک روش آماری تحلیل تشخیصی[2]، شباهت بین مدارک درون هر خوشه بر اساس انواع بافتار هم‌نویسه‌ای بررسی می‌شود.

۲. اهدف پژوهش

در پژوهش حاضر سعی می‌شود با تحلیل انواع بافتارهای هم‌نویسه، به شناخت بافتاری دست یافت که توان بیشتری در تعیین شباهت بین مدارک هم‌نویسه‌دار در درون هر خوشهٔ معنایی دارد.

۳. روش پژوهش

این پژوهش به روش تحلیل محتوای کمّی صورت گرفته است. واحد تحلیل در این پژوهش کلمه است. جامعهٔ پژوهش را مجموعه مدارک حاوی هم‌نویسه‌های انگلیسی تشکیل می‌دهد نمونه‌ی هم‌نویسه‌ها و مدارک در پژوهش حاضر به روش هدفمندِ در دسترس انتخاب شده است. آنگونه که «بلیکی» می‌نویسد، این روش نمونه‌گیری از روش‌های غیراحتمالی است و در مواردی به‌کار گرفته می‌شود که شناسایی یک جمعیت خاص غیرممکن یا بسیار پرهزینه است. کاربرد دیگر این روش در مواردی است که پژوهشگر قصد دارد به انتخاب موارد معدودی از یک نوع خاص بپردازد. این انتخاب بسته به قضاوت پژوهشگر است که کدام موارد برای مقاصد وی مناسب‌ترند (بلیکی، ۱۳۸۴). در این پژوهش شناسایی تمامی هم‌نویسه‌ها و همچنین، تمامی مدارک حاوی هم‌نویسه‌ها

1. test collection
2. discriminant analysis

به‌سادگی امکان‌پذیر نبود. همچنین، بسیاری از هم‌نویسه‌ها عام‌تر از آن بودند که در عنوان مقالات علمی ظاهر شوند. برای برخی نیز پس از بررسی‌های بسیار سرنخ‌های موضوعی دقیقی که بتواند به تمایز معنایی هم‌نویسه کمک کند، یافت نشد. علاوه بر این، درصدی از مدارک شناسایی‌شده از طریق پایگاه‌های علمی آزاد، اشتراکی یا خدمات تحویل مقاله قابل دسترس نبودند. از این رو، در این پژوهش به نمونهٔ هدفمندِ در دسترس بسنده شد. نمونه‌گیری در سه مرحله، شامل انتخاب هم‌نویسه از پژوهش Twilley et al. (1994)، انتخاب معناهای متفاوت هم‌نویسه‌ها و بالاخره، شناسایی مدارک در هر معنا صورت گرفته است.

ساخت مجموعهٔ آزمایشی: برای انجام این پژوهش نیاز بود که در ابتدا یک مجموعهٔ آزمایشی ساخته شود. «سپارک جونز و فن‌ریسبرگن» حداقلِ اندازهٔ ایدئال برای یک مجموعهٔ آزمایشی را ۲۰۰۰ مدرک می‌دانند (Jones & von Rijsbergen 1976 نقل در Harmandas, Sanderson & Dunlop 1997). لازم به ذکر است که در مقالات حوزهٔ بازیابی اطلاعات، کمترین تعداد هم‌نویسهٔ مورد بررسی ۲ و بیشترین تعداد ۲۰ هم‌نویسه برای تحقیق یا آزمایش انتخاب شده (Riahi & Sedghi 2012; Makki & Homayounpour 2008). در این پژوهش، بیشترین تعداد هم‌نویسه‌ها که در متون بازیابی اطلاعات بررسی شده بود، به‌عنوان معیار اولیه در نظر گرفته شد. بررسی هم‌نویسه‌ها تا جایی ادامه پیدا کرد که دست کم یک مجموعهٔ حاوی ۲۰۰۰ مدرک را که توسط (Jones & von Rijsbergen 1976 نقل در Harmandas, Sanderson & Dunlop 1997) به‌عنوان حداقل مطلوب برای یک مجموعهٔ آزمایشی پیشنهاد شده است، محقق سازد.

برای جست‌وجو و یافتن مدارکی که دقیقاً به همان موضوع مورد نظر هم‌نویسه‌ها پرداخته باشد، نیاز بود فرمول‌هایی با توجه به تمامی قابلیت‌ها و تسهیلاتی که در پایگاه «گوگل‌اسکالر» وجود دارد، تهیه شود. سپس، هنگام جست‌وجو، به‌منظور حصول اطمینان از ربط موضوعی مدارک، تمام جست‌وجوها به‌عنوان («allintitle:») محدود شد. همچنین، با این راهبرد، تمامی عناوین حاوی هم‌نویسه مورد جست‌وجو بودند و در نتیجه، بافتار هم‌نویسهٔ عناوین قابل بررسی شد. علاوه بر این، تاریخ جست‌وجو به جدیدترین مدارک (در بازهٔ زمانی ۲۰۰۰-۲۰۱۵) و جست‌وجوها تنها به زبان انگلیسی محدود شدند. بنابراین، احتمال هم‌نویسگی بین زبانی رفع شد. همچنین، جست‌وجو تنها به مقاله‌ها[1] محدود

1. article

شـد و پروانه‌هـای ثبـت اختـراع[1] و اسـتنادها از جسـت‌وجو مسـتثنا شـدند. در نهایـت، بررسـی هم‌نویسـه‌ها بـه انتخـاب تعـداد ۱۹ هم‌نویسـه در ۵۴ گـروه موضوعـی و سـاخت یـک مجموعـه حـاوی ۳۶۳۷ مـدرک منجـر شـد. در ایـن مرحلـه از آنجـا کـه انـدازهٔ مجموعـه را کـه حـدود دو برابـر حداقـل مطلـوب مجموعـهٔ آزمایشـی می‌داننـد (Jones & von Rijsbergen 1976 نقـل در Harmandas, Sanderson & Dunlop 1997)، رسـیده بـود، بررسـی هم‌نویسـه‌ها متوقـف شـد. در جـدول ۱، تعـدادی از هم‌نویسـه‌ها به‌همـراه گـروه موضوعـی خـود نمایـش داده می‌شـود.

جدول ۱. هم‌نویسه‌ها و معناهای مورد بررسی

ردیف	هم‌نویسه	معنا	موضوع	ردیف	هم‌نویسه	معنا	موضوع
۱	Arms	بازو	عضلات بازو	۶	bat	خفاش	شنوایی خفاش
		اسلحه	مسابقات تسلیحاتی			چوب	چوب بیس‌بال
۲	Ball	توپ	شتاب توپ	۷	Cell	یاخته	سلول‌های بنیادی جنینی
		مفصل	مفصل و گوی			موبایل	تلفن همراه و رانندگی
		اجرام یا اشیا (کیهان‌شناسی)	اجرام فرمی			بند	بند زندان
		گوی، گلوله، ساچمه	نانولوله‌ها در آسیاب‌های ساچمه‌ای			خانه (آرایه)	سلول‌های خورشیدی پلیمری
		کره	کرهٔ چشم				
۳	Bank	ساحل	فرسایش ساحل رودخانه	۸	Palm	روغن پالم	روغن پالم در تولید بیودیزل
		بانک داده	بانک داده‌های پروتئین (PDB) (اسم خاص)			کف دست	شناسایی خطوط کف دست
		بانک (بنگاه اقتصادی)	بحران مالی و اعتبار بانکی			سیستم عامل پالم	سیستم عامل پالم
۴	Banking	ذخیره‌سازی	ذخیره‌سازی آب	۹	Press	پرس (بدنسازی)	پرس عضلات سینه
		بانکداری	بانکداری اسلامی			مطبوعات	تبلیغات در مطبوعات

1. patent

ردیف	هم‌نویسه	معنا	موضوع	ردیف	هم‌نویسه	معنا	موضوع
۵	Bass	نوعی ماهی	پرورش ماهی باس	۱۰	Ram	رم (کامپیوتر)	حافظهٔ رم
		نوعی موسیقی	موسیقی باس			قوچ	تولید مثل و جفت‌گیری در قوچ
		نوعی مدل ریاضی	مدل باس			فشار رم (فشار برخوردی تهی‌کننده)	فشار رم در کهکشان‌ها
		اسم خاص	تنگهٔ باس (بین جزیرهٔ تاسمانی و بخش جنوبی استرالیا و ایالت ویکتوریا)			سیلندر	سلندر هیدرولیکی

متن کامل مقالات یافت‌شده به لحاظ انواع بافتار هم‌نویسه‌ها مورد تحلیل قرار گرفت. جدول ۲، انواع بافتار هم‌نویسه‌ها و تعریف آن‌ها را ارائه می‌کند.

جدول ۲. تعریف انواع بافتارهای هم‌نویسه‌ای مورد بررسی

نوع بافتار	تعریف
بافتار هم‌نویسه	واژگان پیرامون هم‌نویسه‌ها در پنجرهٔ واژه‌ای ±۵
بافتار ارجاع	واژگان پیرامون هم‌نویسه‌ها در ارجاع مورد استفاده در مدارک هم‌نویسه‌دار که علاوه بر عنوان می‌تواند کلماتی را از نام نویسنده یا مجله مربوطه دربرگیرد.
بافتار استناد متنی	واژه‌های پیرامون هم‌نویسه‌ها در اطراف استناد درون متنی
بافتار متنی هم‌نویسه	منظور واژگان پیرامون هم‌نویسه‌ها در متن مدارک (به استثنای دیگر بافتارهای مورد بررسی)
بافتار عنوان ارجاع	واژگان پیرامون هم‌نویسه‌ها در عنوان ارجاعات مورد استفاده در مدارک حاوی هم‌نویسه‌ها

با توجه به روندی که در ادبیات به‌کار گرفته شده است (بزم‌آرا و جعفری، بزم‌آرا ۱۳۹۲؛ معروفی و پیله‌ور ۱۳۹۰)، بافتار در فاصلهٔ ۵ واژه قبل و بعد از هم‌نویسه (به‌عبارت دیگر، در پنجره ±۵ واژه) به‌صورت دستی شناسایی و استخراج شدند. در استخراج بافتارها چند نکته رعایت شده است که شرح آن در زیر می‌آید:

۱. سبک‌های استناد در مجلات مختلف یکسان نیست. به همین دلیل، برای ایجاد یکدستی با انواع سبک‌های استنادی یکسان برخورد شده است. یعنی تنها بافتاری که هم‌نویسه در اطراف استناد اتفاق افتاده است، انتخاب شده و خود اطلاعات استناد (شامل نام نویسنده و تاریخ) به شمارش در نیامده است؛

۲. حداکثر طول هر بافتار ۱۰ کلمه در نظر گرفته شد. اما در بافتار عنوان و عنوان ارجاع ممکن بود هم‌نویسه در ناحیه‌ای واقع شده باشد که قبل و بعد از آن کلمه‌ای وجود نداشته یا تعداد کمی از کلمات وجود داشته باشد. در این‌گونه موارد به همان بافتار کوتاه بسنده شد؛
۳. از انتخاب بافتار پانویس‌ها و جدول‌ها صرف نظر شد؛
۴. حروف اضافه و تعریف از بافتار حذف نشدند و این رویکرد در انتخاب تمامی بافتارها رعایت شده است. این رویکرد از آن رو اتخاذ شد که در سامانه‌های جدید بازیابی اطلاعات تمایل به عدم حذف واژگان بازدارنده[1] وجود دارد (Manning, Raghavan & Schütze 2008). علاوه بر این، از آنجا که در محاسبهٔ وزن کلمات از فرمول به هنجارشده tf-idf استفاده می‌شود، فراوانی این کلمات در مقادیر این شاخص خنثی می‌شود؛
۵. کلمات مرکب و کلماتی که دارای خط تیره بودند، در میان داده‌ها اندک بود. ولی، به‌منظور رعایت یکدستی این‌گونه کلمات یک کلمهٔ واحد به حساب آمدند.

در مجموع، تحلیل محتوای این مدارک بازیابی‌شده به شناسایی ۲۶۰۰۳ واژهٔ منحصربه‌فرد انجامید که در مجموع، ۹۰۶۶۵۷ بار در ۵ نوع بافتار روی داده بودند. این کلمات برای آماده‌سازی به EXCEL منتقل شد. به کمک الگوریتم «پرتر»[2]، پایانه‌های صرفی، وندها و همچنین علائم سجاوندی، اِعراب و جز آن حذف شد. شایان ذکر است که پیش از به‌کارگیری ریشه‌یاب «پرتر»، اعداد اعشاری نیز به‌طور دستی بهنجار شد تا علامت ممیز انگلیسی با نقطهٔ سجاوندی اشتباه گرفته نشود. سپس، برای تجزیه و تحلیل آن‌ها از نرم‌افزار SPSS و از روش آماری تحلیل تشخیصی استفاده شد.

روش محاسبهٔ ارزش و وزن واژه‌ها: در نخستین گام تحلیل داده‌ها، با اختصاص هر واژه به مدرک مربوطه، ماتریس رویداد واژه–مدرک تشکیل شد. آن‌گاه به هر واژه به روشی که در پی می‌آید، وزن داده شد.

نخست، به هر واژه در هر مدرک بر اساس شاخص"tf-idf" وزن داده شد. مقدار"tf-idf" برای هر واژه بر اساس فرمول به هنجارشده زیر محاسبه شد:

$$\mathbf{tf-idf} = tf \times idf$$

1. stop words
2. Porter

که در آن:
tf= فراوانی واژه در مدرک؛ *Idf*= وارون فراوانی مدارکی که واژه در آن روی داده است؛ مقدار *idf* بر اساس فرمول زیر به هنجار شد:

$$idf = log \frac{N}{df}$$

که در آن *log*= لگاریتم در مبنای ۱۰؛ *N* = تعداد کل مدارک در مجموعه؛ *df* = فراوانی مدارکی که واژه در آن روی داده است (Manning, Raghavan & Schütze 2008).

این مقدار در واقع، ارزش هر واژه را در کل یک مدرک به‌دست می‌دهد. با توجه به آن که طول هر مدرک می‌تواند بر مقدار *tf* آن تأثیر بگذارد، مقدار *tf* نیز بر اساس مقدار بیشینه *tf* هر مدرک به‌هنجار می‌شود:

$$ntf_{d,t} = a + (1 - a)\frac{tf_{t,d}}{tf_{\max(d)}}$$

که در آن a مقداری است بین صفر و یک و معمولاً برابر با ۰/۴ در نظر گرفته می‌شود (همان). در این پژوهش نیز طول بافتارها و همچنین، طول مدارک با هم متفاوت بود و این می‌توانست بر مقدار *tf* تأثیر بگذارد. از این رو، برای به‌هنجارسازی طول مدرک از این فرمول استفاده شد.

از آنجا که نمرهٔ هر مدرک برای واژگان روی داده در هر یک از بافتارهای آن به‌طور جداگانه محاسبه شد، ۵ نوع نمرهٔ متفاوت برای هر مدرک به‌دست آمد که به ۵ بافتار عنوان، استناد، متن، ارجاع و عنوان ارجاع مرتبط بود.

روش تجزیه و تحلیل داده‌ها

در این پژوهش از روش تحلیل تشخیصی گام‌به‌گام[1] استفاده شد. تحلیل تشخیصی، شناسایی متغیرهای تمیزدهنده را در دو مرحله انجام می‌دهد: در مرحلهٔ اول، از آزمون F (و بر حسب مقدار لاندای ویلکس) برای آزمون معناداری کل مدل تشخیصی استفاده می‌شود. در مرحلهٔ دوم، چنانچه مقدار F در مرحلهٔ اول معناداری مدل تشخیصی را نشان دهد، در آن صورت هر متغیر مستقل ارزیابی می‌شود تا مشخص شود میانگین کدام‌یک از آن‌ها تفاوت معناداری با همدیگر بر حسب گروه دارد که بتوان در نهایت، از آن‌ها برای طبقه‌بندی متغیر وابسته استفاده کرد (حبیب‌پور گتابی و صفری شالی ۱۳۸۸). همچنین،

1. stepwise discriminant analysis

در روش گام‌به‌گام، نخست قوی‌ترین متغیر پیش‌بین انتخاب می‌شود، واریانس در متغیر گروه‌بندی (متغیر ملاک) حذف می‌شود و سپس، قوی‌ترین متغیر پیش‌بین بعدی به تحلیل افزوده می‌شود و این کار تا زمانی که تغییرات همبستگی کانونی معنادار باشد، ادامه می‌یابد.

پیش از انجام این تحلیل، تمامی پیش‌فرض‌ها شامل نرمالیتی، همگن‌بودن و همبسته‌نبودن متغیرها کنترل شد.

در تحلیل تشخیصی از معیارهای متعددی جهت آزمون معناداری تمایز و همچنین، مقایسهٔ توابع به‌دست‌آمده استفاده می‌شود. مقدار ویژه[1]، نسبت بین واریانس میان‌گروهی به درون‌گروهی است. هرچه این مقدار بیشتر باشد، نشان از قوت آماری تحلیل تشخیصی دارد.

درصد واریانس[2]، بخشی از واریانس تحلیل تشخیصی را که توسط متغیرهای واردشده به تحقیق تبیین می‌شود، نشان می‌دهد.

همبستگی کانونی[3]، همبستگی میان گروه‌ها و تابع تشخیصی را نشان می‌دهد و بر اساس ریشهٔ دوم نسبت بین واریانس میان‌گروهی و واریانس کل محاسبه می‌شود و در آخر لاندای ویلکس[4] است که عبارت است از نسبت مجموع مجذورات درون‌گروهی به مجموع مجذورات کل. در واقع، لاندای ویلکس نسبتی از واریانس کل نمرات تشخیصی است که توسط تفاوت بین گروه‌ها تبیین نشده است. بنابراین، هرچه مقدار لاندای ویلکس کمتر باشد، بهتر و قدرت تبیین‌گری مدل بیشتر است. مقدار این آماره بین (۰) تا (۱) نوسان دارد و بر اساس فرمول زیر محاسبه می‌شود:

$$(1-R_1^2)^* (1-R_2^2)$$

که در آن R برابر است با ضریب همبستگی (ضریب تعیین) میان هر متغیر و تابع تشخیصی. هرچه مقادیر معیارهای مقدار ویژه، درصد واریانس و همبستگی کانونی برای تابعی بالاتر باشد و در مقابل مقدار لاندای ویلکس برای آن تابع پایین‌تر باشد، آن تابع از قوت آماری بیشتری برخوردار است.

شایان ذکر است که در تحلیل تشخیصی مقدار پیش‌گزیده برای F در اس‌پی‌اس‌اس (F مجاز) انتخاب شد. مقدار F پیش‌گزیده برای ورود به تحلیل برابر است با ۳/۸۴ و مقدار

1. Eigen value 2. % of Variance 3. Canonical Correlation 4. Wilks' Lambda

F پیش‌گزیده برای حذف از تحلیل برابر است با ۲/۷۱. روشن است که هرچه مقدار F مجاز کوچک‌تر باشد، گام‌های تحلیل بیشتر شده و دقت نتایج خوشه‌بندی افزایش می‌یابد.

٤. تجزیه و تحلیل یافته‌ها

برای پاسخگویی به پرسش‌های پژوهش با استفاده از تحلیل تشخیصی، تمایز میان مدارک در درون هر یک از خوشه‌های معنایی بررسی شد. مدارک به‌عنوان زیرخوشه‌ها در درون هر خوشهٔ معنایی در نظر گرفته شدند. هدف از تحلیل‌های درون‌خوشه‌ای شناسایی بافتارهایی بود که کمترین نقش را در ساخت این زیرخوشه‌ها و در نتیجه، بیشترین شباهت را در میان مدارک داشته‌اند. این تحلیل‌ها در سطح واژه انجام شد. جهت سهولت در خوانش جداول از اختصارات CIT برای بافتار «استناد»، REF برای بافتار «ارجاع»، REF_TI برای بافتار «عنوان ارجاع»، TI برای بافتار «عنوان مقاله»، و بالاخره TXT برای بافتار «متن» استفاده شده است.

چنان‌که در جدول ۳، مشاهده می‌شود، نتایج تحلیل تشخیصی در درون ۱۳ خوشه از ۵۴ خوشهٔ موضوعی مورد بررسی معنادار نبوده است. این خوشه‌ها عبارت‌اند از: چهار خوشهٔ "eye ball"، "ball Fermi"، "ball mill" و "ball velocity"، خوشهٔ "protein data bank" از بین سه خوشهٔ هم‌نویسهٔ "Bank"، دو خوشهٔ "bass model" و "sea bass" از بین چهار خوشهٔ هم‌نویسهٔ "Bass"، سه خوشهٔ "log linear"، "query log" و "saw log" در هم‌نویسهٔ "Log"، خوشهٔ "pupil mathematics" از بین دو خوشه در هم‌نویسهٔ "Pupil"، خوشهٔ "ram pressure" از بین سه خوشه در هم‌نویسه "Ram"، خوشهٔ "Frozen elephant trunk" از بین چهار خوشه در هم‌نویسهٔ "Trunk". این بدان معناست که مدارک موجود در درون این خوشه‌ها به‌لحاظ انواع بافتارها تفاوت معناداری با هم نداشته‌اند. به‌عبارت دیگر، همهٔ مدارک در درون این خوشه‌ها به لحاظ متغیرهای مورد بررسی شبیه بوده‌اند و بنابراین، انواع پنج‌گانهٔ بافتار متن نتوانسته‌اند بین مدارک موجود در درون این خوشه‌ها تمایزی ایجاد کنند.

چنان‌که در جدول ۳، مشاهده می‌شود، نتایج تحلیل تشخیصی در درون هر خوشه نشانگر آن است که در ۲۳ خوشه تنها یک نوع بافتار به نحوی معنادار بین مدارک درون خوشه‌ها متفاوت بوده است و ۴ نوع بافتار دیگر به نحوی به هم شبیه بوده‌اند که نتوانسته‌اند در ایجاد تمایز بین مدارک در درون خوشه‌ها نقشی داشته باشند. خوشهٔ "bed

"bug" تنها خوشه‌ای است که در آن بافتار متن تنها بافتار متمایز بین مدارک بوده است (sig.=۰/۰۰۰، F=۵/۱۲۴ و λ=۰/۹۶۶). در دو خوشهٔ "ball joint" (sig.=۰/۰۰۰، F=۱۸/۸۷۵ و λ=۰/۹۴۰) و "eye pupil" (sig.=۰/۰۰۰، F=۵/۹۶۸ و λ=۰/۹۶۱) بافتار ارجاع تنها بافتاری است که در تمایز بین مدارک در درون هر خوشه نقش داشته است. این بدان معناست که مدارک در درون هر خوشه به لحاظ چهار بافتار دیگر یعنی بافتار استناد، عنوان ارجاع، متن و عنوان مقاله شباهت داشته‌اند. بافتار عنوان ارجاع در درون هر یک از سه خوشهٔ "bat auditory" (sig.=۰/۰۰۰، F=۴/۳۸۱ و λ=۰/۹۷۹)، "stem cell" (sig.=۰/۰۰۰، F=۳/۹۵۹ و λ=۰/۹۷۸) و "ram breeding" (sig.=۰/۰۰۰، F=۳/۹۰۸ و λ=۰/۹۶۶) به تمایز میان مدارک انجامیده است (جدول ۳).

در ۱۷ خوشه از این ۲۳ خوشه، بافتارهای عنوان مدارک در درون هر خوشه با هم به‌گونه‌ای معنادار متفاوت هستند، به نحوی که به ایجاد زیرخوشه‌هایی در درون هر خوشه انجامیده‌اند. این خوشه‌ها عبارت‌اند از: خوشهٔ "arms muscle" (sig.=۰/۰۰۰، F=۵/۰۰۵ و λ=۰/۹۵۰)، خوشهٔ "bank river" (sig.=۰/۰۰۰، F=۷/۱۵۶ و λ=۰/۹۵۴)، خوشهٔ "Islamic banking" (sig.=۰/۰۰۰، F=۸/۰۹۹ و λ=۰/۹۲۴)، خوشهٔ "bat baseball" (sig.=۰/۰۰۰، F=۸/۳۰۷ و λ=۰/۹۴۱)، خوشهٔ "copper board" (sig.=۰/۰۰۰، F=۶/۱۸۱ و λ=۰/۸۷۴)، خوشهٔ "editorial board" (sig.=۰/۰۰۰، F=۶/۷۲۳ و λ=۰/۹۳۶)، خوشهٔ "ship board" (sig.=۰/۰۰۰، F=۹/۰۰۱ و λ=۰/۸۴۶)، خوشهٔ "bug web" (sig.=۰/۰۰۰، F=۱۷/۵۴۰ و λ=۰/۸۲۵)، خوشهٔ "cell phone" (sig.=۰/۰۰۰، F=۷/۴۰۱ و λ=۰/۹۵۷) و خوشهٔ "solar cell" (sig.=۰/۰۰۰، F=۵/۱۵۱ و λ=۰/۹۴۴)، خوشهٔ "mole cricket" (sig.=۰/۰۰۰، F=۹/۰۳۹ و λ=۰/۹۰۹)، خوشهٔ "palm oil" (sig.=۰/۰۰۰، F=۴/۸۵۳ و λ=۰/۹۳۵)، خوشهٔ "palm print" (sig.=۰/۰۰۰، F=۴/۶۸۲ و λ=۰/۹۶۳)، خوشهٔ "bench press" (sig.=۰/۰۰۰، F=۴/۰۱۳ و λ=۰/۹۴۸)، خوشهٔ "trunk muscles" (sig.=۰/۰۰۰، F=۴/۲۷۳ و λ=۰/۹۵۰) و خوشهٔ "tree trunk" (sig.=۰/۰۰۰، F=۷۵۴/۰۱۸ و λ=۰/۹۲۰).

در ۱۸ خوشهٔ موضوعی باقی‌مانده بیش از یک بافتار باعث تمایز بین مدارک در درون هر خوشه شده است. در این میان، در غالب موارد، بافتار عنوان در کنار دیگر بافتارها به چشم می‌خورد. تنها استثنا در این باره عبارت‌اند از: سه خوشهٔ "bank credit" (sig.=۰/۰۰۰، F=۶/۱۰۲ و λ=۰/۷۳۷)، "hydraulic ram" (sig.=۰/۰۰۰، F=۸/۲۰۲ و λ=۰/۹۵۶) و "computer terminal" (sig.=۰/۰۰۰، F=۹/۰۶۲ و λ=۰/۷۸۶) که در آن‌ها بافتار عنوان بین مدارک در درون هر خوشه شبیه بوده است. در همهٔ خوشه‌های موضوعی دیگر بافتار عنوان یکی از

بافتارهایی است که به تمایز میان مدارک در درون هر خوشه منجر شده است (جدول ۳). برای مثال، در خوشهٔ "arms racing" دو نوع بافتار متن و ارجاع در تمایز میان مدارک در درون این خوشه مؤثر بوده‌اند (sig.=۰/۰۰۰، F=۶/۵۰۱ و λ=۰/۸۲۲). به‌عنوان مثالی دیگر، در خوشهٔ "water banking" دو بافتار متن و عنوان مدارک درون این خوشه را از هم متمایز کرده است (sig.=۰/۰۰۰، F=۶/۶۴۹ و λ=۰/۸۱۹).

با آن‌که تمایزی معنادار بین زیرخوشه‌های درون خوشه‌های معنایی مشاهده شده است، اما مقادیر لاندای ویلکس به‌دست‌آمده عمدتاً بالاست. همان‌گونه که در جدول و در شرح بالا ملاحظه می‌شود، به‌جز شش خوشهٔ "bass music" (۰/۵۶۹)، "bass strait" (۰/۶۸۷)، "prison cell" (۰/۶۲۶)، "ball joint" (۰/۷۳۷)، "computer terminal" (۰/۷۸۶) و "cord terminal" (۰/۷۱۲)، در دیگر موارد مقادیر لاندای ویلکس بالاتر از ۰/۸ شده است. این بدان معناست که تمایز بین مدارک در درون خوشه‌ها بسیار ناچیز است و در واقع، می‌توان شباهت همهٔ بافتارها را به یک اندازه برآورد کرد[1]. این بدان معناست که بافتارهای مورد بررسی تمایز بسیار بسیار ناچیزی در میان مدارک در درون خوشه‌ها ایجاد کرده‌اند و عملاً بین بافتارها به لحاظ قدرت تعیین شباهت مدارک در درون خوشه‌های معنایی تفاوت چشمگیری دیده نمی‌شود. با این حال و بر مبنای همان مقدار تمایز ناچیز، به نظر می‌رسد که بافتار عنوان از کمترین قدرت در تعیین شباهت در میان مدارک در درون خوشه‌ها برخوردار بوده است و چهار بافتار دیگر کم‌وبیش قوت یکسانی در تعیین شباهت معنایی درون خوشه‌ها داشته‌اند.

۱. شایان ذکر است که در این مرحله از پژوهش، به‌منظور بررسی متغیرهایی که بیشترین شباهت را در درون هر یک از ۵۴ خوشهٔ معنایی داشتند، از تحلیل خوشه‌ای دو مرحله‌ای نیز استفاده شد. در نظر بود به کمک این آزمون زیرخوشه‌های احتمالی در درون هر خوشه شناسایی شود و بافتاری که بیشترین نقش را در ساخت این زیرخوشه‌ها داشته است، تعیین گردد. نتایج همهٔ تحلیل‌ها، (به‌جز دو مورد) نشان داد که هیچ‌گونه زیرخوشه‌ای در درون هر خوشهٔ معنایی وجود ندارد و تحلیل به یک خوشهٔ واحد برای هر یک از خوشه‌های معنایی دست یافت. در نتیجه، این تحلیل‌ها نیز نتایج تحلیل تشخیصی درون خوشه‌ای را مبنی بر ناچیزبودن تمایز بین مدارک در درون خوشه‌های معنایی و شباهت همهٔ بافتارها در میان مدارک در درون خوشه‌ها را تأیید می‌کند. با این حال، به جهت رعایت اختصار از ارائهٔ نتایج آن تحلیل‌ها خودداری می‌شود.

جدول ۳. تحلیل تشخیصی در درون خوشه‌ها

ردیف	همنویسه	موضوع	متغیرهای واردشده به تحلیل	متغیرهای واردنشده به تحلیل	لاندای ویلکس	آمارZ آزمون	سطح معناداری
۱	Arms	arms muscle	TI	REF, CIT, TXT, REF_TI	۰/۹۵۰	۵/۰۰۵	۰/۰۰۰
		arms racing	TI, REF	CIT, TXT, REF_TI	۰/۸۲۲	۶/۵۰۱	۰/۰۰۰
۲	Ball	eye ball	NS				
		ball Fermi					
		ball joint	REF	TI, CIT, TXT, REF_TI	۰/۷۳۷	۱۸/۸۷۵	۰/۰۰۰
		ball mill	NS				
		ball velocity					
۳	Bank	bank credit	REF, TXT, REF_TI	TI, CIT	۰/۹۴۰	۶/۱۰۲	۰/۰۰۰
		protein data bank	NS				
		bank river	TI	REF, CIT, TXT, REF_TI	۰/۹۵۴	۷/۱۵۶	۰/۰۰۰
۴	Banking	Islamic banking	TI	REF, CIT, TXT, REF_TI	۰/۹۲۴	۸/۰۹۹	۰/۰۰۰
		water banking	TI, TXT	REF, CIT, REF_TI	۰/۸۱۹	۶/۶۴۹	۰/۰۰۰
۵	Bass	bass model	NS				
		bass music	TI, TXT, REF_TI	REF, CIT	۰/۵۶۹	۱۶/۴۳۶	۰/۰۰۰
		sea bass	NS				
		bass strait	TI, TXT, REF_TI	REF, CIT	۰/۶۸۷	۷/۷۹۴	۰/۰۰۰
۶	Bat	bat auditory	REF_TI	TI, REF, CIT, TXT	۰/۹۷۹	۴/۳۸۱	۰/۰۰۰
		baseball bat	TI	REF, CIT, TXT, REF_TI	۰/۹۴۱	۸/۳۰۷	۰/۰۰۰
۷	Board	copper board	TI	REF, CIT, TXT, REF_TI	۰/۸۷۴	۶/۱۸۱	۰/۰۰۰
		editorial board	TI	REF, CIT, TXT, REF_TI	۰/۹۳۶	۶/۷۲۳	۰/۰۰۰
		ship board	TI	REF, CIT, TXT, REF_TI	۰/۸۴۶	۹/۰۰۱	۰/۰۰۰

ردیف	همنویسه	موضوع	متغیرهای وابسته به تحلیل	متغیرهای واردنشده به تحلیل	لاندای ویلکس	آمارZ آزمون	سطح معناداری
۸	Bug	bed bug	TXT	TI, REF, CIT, REF_TI	۰/۹۶۶	۵/۱۲۴	۰/۰۰۰
		bug web	TI	REF, CIT, TXT, REF_TI	۰/۸۲۵	۱۷/۵۴۰	۰/۰۰۰
۹	Cell	cell phone	TI	REF, CIT, TXT, REF_TI	۰/۹۵۷	۷/۴۰۱	۰/۰۰۰
		prison cell	TI, CIT, TXT, REF_TI	REF	۰/۶۲۶	۱۲/۴۸۹	۰/۰۰۰
		solar cell	TI	REF, CIT, TXT, REF_TI	۰/۹۴۴	۵/۱۵۱	۰/۰۰۰
		stem cell	REF_TI	TI, REF, CIT, TXT	۰/۹۷۸	۳/۹۵۹	۰/۰۰۰
۱۰	Cricket	mole cricket	TI	REF, CIT, TXT, REF_TI	۰/۹۰۹	۹/۰۳۹	۰/۰۰۰
		cricket bat	TI, TXT, REF_TI	REF, CIT	۰/۸۳۹	۶/۷۰۱	۰/۰۰۰
۱۱	Log	log linear	NS				
		query log					
		saw log					
۱۲	Matches	exact matches	TI, TXT, REF_TI	REF, CIT	۰/۸۰۱	۶/۱۳۸	۰/۰۰۰
		football matches	TI, REF	CIT, TXT, REF_TI	۰/۸۸۳	۵/۹۷۹	۰/۰۰۰
۱۳	Palm	palm oil	TI	REF, CIT, TXT, REF_TI	۰/۹۳۵	۴/۸۵۳	۰/۰۰۰
		palm OS	TI, REF	CIT, TXT, REF_TI	۰/۸۹۴	۶/۰۸۷	۰/۰۰۰
		palm print	TI	REF, CIT, TXT, REF_TI	۰/۹۶۳	۴/۶۸۲	۰/۰۰۰
۱۴	Press	Press advertisements	TI, REF	CIT, TXT, REF_TI	۰/۹۶۹	۹/۵۷۸	۰/۰۰۰
		bench press	TI	REF, CIT, TXT, REF_TI	۰/۹۴۸	۴/۰۱۳	۰/۰۰۰
۱۵	Pupil	eye pupil	REF	TI, CIT, TXT, REF_TI	۰/۹۶۱	۵/۹۶۸	۰/۰۰۰
		pupil mathematics	NS				

ردیف	همنویسه	موضوع	متغیرهای واردشده به تحلیل	متغیرهای واردنشده به تحلیل	لانـدای ویلکس	آمارZ آزمون	سطح معناداری
۱۶	Ram	ram breeding	REF_TI	TI, REF, CIT, TXT	۰/۹۶۶	۳/۹۰۸	۰/۰۰۰
		hydraulic ram	REF, TXT, REF_TI	TI, CIT	۰/۹۵۶	۸/۲۰۲	۰/۰۰۰
		ram memory	TI, REF_TI	REF, CIT, TXT	۰/۸۷۵	۶/۵۷۶	۰/۰۰۰
		ram pressure	NS				
۱۷	Terminal	bus terminal	TI, TXT	REF, CIT, REF_TI	۰/۸۶۷	۵/۹۰۹	۰/۰۰۰
		computer terminal	TXT, REF_TI	TI, REF, CIT	۰/۷۸۶	۹/۰۶۲	۰/۰۰۰
		cord terminal	TI, REF_TI, CIT	REF, TXT	۰/۷۱۲	۸/۶۸۸	۰/۰۰۰
		patients terminal	TI, TXT	REF, CIT, REF_TI	۰/۸۴۴	۶/۱۶۷	۰/۰۰۰
۱۸	Trunk	Frozen elephant trunk	NS				
		trunk muscles	TI	REF, CIT, TXT, REF_TI	۰/۹۵	۴/۲۷۳	۰/۰۰۰
		tree trunk	TI	REF, CIT, TXT, REF_TI	۰/۹۲	۷۵۴/۰۱۸	۰/۰۰۰
۱۹	Web	android web	TI, REF, CIT	TXT, REF_TI	۰/۸۳۴	۴/۸۵۷	۰/۰۰۰
		web building	TI, TXT, REF_TI	REF, CIT	۰/۸۶۶	۶/۱۱۶	۰/۰۰۰

نگاهی به فراوانی بافتارهایی که قدرت تمایزگری بالایی داشته‌اند، به درک بهتر این واقعیت کمک می‌کند. به‌طور کلی، همان‌گونه که در جدول مشاهده می‌شود، بافتار استناد تنها در ۳ مورد به تمایز بین مدارک در درون خوشه‌ها منجر شده است. این بدان معناست که در ۵۱ خوشهٔ دیگر بافتار استناد به هم شبیه بوده است. دو بافتار عنوان ارجاع و ارجاع نیز هر یک به‌ترتیب، با ۱۱ و ۱۶ فراوانی در حد میانه قرار دارند. به عبارت دیگر، این دو بافتار نیز در غالب موارد (به‌ترتیب ۴۳ و ۳۸ خوشه) در میان مدارک درون هر خوشه شبیه بوده‌اند. در مقابل، بافتار عنوان در ۴۶ مورد به تمایز میان مدارک در درون خوشه‌ها انجامیده است و تنها در ۸ خوشه بافتار عنوان مشابه بوده است.

۵. بحث و نتیجه‌گیری

پیش‌فرض این پژوهش آن بود که هم‌معنا بودن هم‌نویسه‌ها به ایجاد خوشه‌هایی

از مدارک منجر می‌شود که در درون آن‌ها مدارک با هم به لحاظ انواع بافتار متنی تفاوت معناداری ندارند. از این رو، با شناسایی خوشه‌هایی از هم‌نویسه‌های دارای یک معنای واحد تلاش شد تا رده‌هایی از قبل تعریف شود و به کمک تحلیل تشخیصی مورد بررسی قرار گیرد. بر اساس این پیش‌فرض، انتظار می‌رفت که در درون خوشه‌های معنایی زیرخوشه‌هایی به لحاظ انواع بافتار مشاهده نشود. نتایج این پژوهش نشان داد که تمایز معناداری میان مدارک در درون ۱۳ خوشه هم‌معنا وجود ندارد. ۴۱ خوشهٔ معنایی دیگر، در درون خود دارای زیرخوشه‌هایی از مدارک هستند که به لحاظ انواع بافتار به‌ویژه بافتار عنوان با هم متفاوت هستند. با این حال، مقدار لاندای ویلکس برای تحلیل‌های درون خوشه‌ها در بیشتر موارد بسیار بالا بود و نشان از عدم قدرت پیش‌بینی بالای تمایز بین خوشه‌ها داشت. در نتیجه، علی‌رغم معناداربودن تمایز بین خوشه‌ها، این تمایز بسیار ناچیز و شباهت بین بافتارهای متنی در مدارک در درون خوشه‌ها بالا برآورد می‌شود. نتایج حاصل از این بخش مبنی بر شباهت میان مدارک در درون هر خوشهٔ معنایی به لحاظ بافتارهای متنی با یافته‌های حاصل از پژوهش‌های Church & Hanks (1990) همراستاست.

تا جایی که بررسی‌های به‌عمل‌آمده در متون نشان داد، هیچ‌گونه پژوهشی یافت نشد که به مقایسهٔ قدرت انواع بافتار متن در تعیین شباهت میان مدارک پرداخته باشد. به این ترتیب، آنچه نتایج این پژوهش به دانش پیشین می‌افزاید، علاوه بر تأکید دوباره بر نقش بافتارهای متنی پیرامون کلمه در ابهام‌زدایی معنایی از مدارک حاوی آن‌ها، تأیید تفاوت در انواع بافتارها به لحاظ قدرت پیش‌بینی شباهت بین مدارک است؛ به‌نحوی که بافتار استناد متنی و بافتار ارجاع (اندکی) بیش از بافتارهای متن، عنوان ارجاع یا عنوان مقاله در تعیین شباهت معنایی بین مدارک اثرگذار هستند.

بافتارهای مورد بررسی در این پژوهش را می‌توان به لحاظ بنیان‌های نظری به دو گروه تقسیم کرد. دو بافتار عنوان و متن در واقع، ارتباط تنگاتنگی با محتوای آثار علمی دارند و «دربارگی نویسنده» را بازتاب می‌دهند. مفهوم موضوعیت یا دربارگی، اساساً به این اشاره دارد که یک مدرک، متن، تصویر و جز آن دربارهٔ «چیست» و «چه‌کسی» دربارهٔ این «چیستی» تصمیم می‌گیرد. نویسنده از طریق بازنمون زبان طبیعی «موضوعیت» اثر را تعیین می‌کند (اینگورسن ۱۹۹۲) و آن را در متن خود منعکس می‌سازد. به واقع، پیش‌فرض آن است که نویسندهٔ متخصص به بهترین وجه با مطلب در دست بررسی خود آشناست و از

بهترین و مرتبط‌ترین واژگان در اثر خود استفاده می‌کند. این موضوعیت به بهترین وجه در بافتار متن و بافتار عنوان منعکس می‌شود. سه بافتار دیگر این تحقیق یعنی بافتار استناد متنی، بافتار ارجاع و بافتار عنوان ارجاع به‌طور زیربنایی به نظریه‌های استنادی مرتبط می‌شود. بافتار استناد متنی عبارت است از متنی که نویسنده در پیرامون عبارت استنادی خود به‌کار برده است. بافتار عنوان ارجاع به عنوان آثار مورد استناد نویسنده اشاره دارد و تفاوت آن با بافتار ارجاع در آن است که بافتار ارجاع علاوه بر عنوان می‌تواند کلماتی را از نام نویسنده یا مجله مربوطه دربرگیرد.

یکی از پیش‌فرض‌های تحلیل استنادی این است که بین دو مدرک استنادشونده و استنادکننده ربط موضوعی وجود دارد. بر این اساس، همگی این عوامل اعم از کلمات و عباراتی که نویسنده در متن خود به‌کار می‌برد، عنوان آثار مورد استناد وی، نویسندگان مورد استناد وی، و همچنین، مجلات منتشرکنندهٔ این آثار می‌توانند به‌نحوی با محتوای اثر استنادکننده مرتبط باشند.

آشکار است که این سه بافتار به لحاظ بنیان‌های نظری با دو بافتار دیگر بی‌ارتباط نیستند. زیرا در این سه بافتار نیز این نویسنده است که تصمیم می‌گیرد به چه اثری ارجاع دهد و با چه عبارت‌هایی، مفاهیم برگرفته از تحقیقات مورد استناد خود را تبیین کند. در این میان، بافتار استناد متنی به واقع، آمیزه‌ای از دربارگی نویسندهٔ استنادکننده و نویسندهٔ استنادشونده است، در حالی که دو بافتار ارجاع و عنوان ارجاع بیشتر دربارگی نویسندهٔ استنادشونده را نشان می‌دهد تا نویسندهٔ استنادکننده. زیرا در این دو بافتار اخیر نویسندهٔ استنادکننده دربارگی خود را تنها در مرحلهٔ انتخاب اثر برای استناد دخالت می‌دهد و نقشی در انتخاب واژگان و تدوین عبارت‌های مربوطه ندارد زیرا در این دو بافتار اخیر نویسندهٔ استنادکننده دربارگی خود را تنها در مرحلهٔ انتخاب اثر برای استناد دخالت می‌دهد و نقشی در انتخاب واژگان و تدوین عبارت‌های مربوطه ندارد. در این تحقیق، علاوه بر ارائهٔ شواهدی تازه که پیش‌فرض ربط موضوعی بین اثر استنادکننده و استنادشونده را تأیید می‌کند، بر قدرت بافتارهای استنادی در تعیین شباهت بین مدارک تاکید شد.

همچنین، آن بخش از نتایج مربوط به اهمیت بافتار عنوان ارجاع و بافتار استناد متنی و بافتار عنوان تنها به لحاظ اهمیت این سه نوع بافتار در خوشه‌بندی مدارک با پژوهش Tong, Dinakarpandian, & Lee (2009) همراستاست. علاوه بر این، آن بخش از نتایج پژوهش

کـه مربـوط بـه اهمیـت بافتـار اسـتناد در خوشـه‌بندی مـدارک هم‌معناسـت، بـا پژوهـش Tsai, Kundu, & Roth (2013) همراستاسـت. آن بخـش از یافته‌هـای ایـن پژوهـش کـه مربـوط بـه قـوت بافتـار اسـتناد اسـت، یافته‌هـای پژوهش‌هـای پیشـین مبنـی بـر قـوت ایـن بافتـار را تأییـد می‌کنـد.

محققـان حـوزهٔ بازیابـی اطلاعـات تـلاش می‌کننـد بـه کمـک ویژگی‌هـا و نشـانه‌های موجـود در خـود متـن بـه درک معنـای محتـوای آن، تعییـن شـباهت آن بـا دیگـر مـدارک یـا پرسـش‌ها و همچنیـن، رتبه‌بنـدی مـدارک نایـل آینـد. از آنجـا کـه متـن می‌توانـد دارای عناصـر بسـیار طولانـی باشـد، شناسـایی بخشـی از اثـر کـه بیـش از همـه در تعییـن ایـن شـباهت بیـن مـدارک نقـش داشـته اسـت، می‌توانـد در افزایـش کارایـی سـامانه مؤثـر باشـد. زیـرا اخص‌سـازی نـوع بافتـار واژگانـی می‌توانـد بخـش اعظمـی از بـار سـامانه را کاسـته و ابهام‌زدایـی معنایـی را بـرای الگوریتم‌هـا سـاده‌تر سـازد. هرچـه ویژگی‌هـا و شـاخص‌های متنـی بیشـتری کشـف و بررسـی گـردد، دسترسـی بـه شـاخص یـا دسـته شـاخص‌هایی کـه در تعییـن شـباهت بیـن مـدارک هم‌نویسـه‌دار بـا معنـای مشـابه از تـوان بیشـتر برخـوردار باشـند، امکان‌پذیرتـر خواهـد شـد. نتایـج حاصـل از پژوهـش حاضـر نشـان داد کـه نه‌تنهـا بافتارهـای متنـی پیرامـون کلمـه در ابهام‌زدایـی معنایـی نقـش ویـژه‌ای داشـته‌اند، بلکـه انـواع بافتارهـا بـه لحـاظ قـدرت پیش‌بینـی شـباهت بیـن مـدارک تفـاوت دارنـد؛ به‌نحـوی کـه بافتـار اسـتناد متنـی و بافتـار ارجـاع تـوان بیشـتری نسـبت بـه بافتارهـای متـن، عنـوان ارجـاع و عنـوان مقالـه در تعییـن شـباهت معنایـی بیـن مـدارک دارنـد. بنابرایـن، هم‌معنـا بـودن هم‌نویسـه‌ها بـه ایجـاد خوشـه‌هایی از مـدارک منجـر می‌شـود کـه در درون آن‌هـا مـدارک بـا هـم بـه لحـاظ انـواع بافتـار تفـاوت معنـاداری ندارنـد.

٦. پیشنهادهای پژوهش

از نتایـج ایـن پژوهـش می‌تـوان بـرای بهبـود نتایـج بازیابـی کمـک گرفـت. ایـن یافته‌هـا می‌توانـد در الگوریتم‌هـای بازیابـی موتورهـای جسـت‌وجو و پایگاه‌هـای اطلاعاتـی بـرای سـنجش دقیق‌تـر شـباهت مـدارک هم‌نویسـه‌دار بـا یکدیگـر مـورد اسـتفاده قـرار گیرنـد. یکـی از دغدغه‌هـای مهـم طراحـان سـامانه‌های بازیابـی اطلاعـات، تضمیـن کارایـی سـامانه در عیـن ارتقـای اثربخشـی آن‌هاسـت. نتایـج پژوهـش حاضـر می‌توانـد در هـر دو جهـت مؤثـر باشـد. از یک‌سـو، بـا تمرکـز بـر انواعـی از بافتـار کـه تـوان بیشـتری در تعییـن شـباهت متـن دارنـد، سـامانه از پـردازش کل متـن بـرای تعییـن شـباهت آن بی‌نیـاز خواهـد شـد و ایـن امـر می‌توانـد بـه کارایـی سـامانه کمـک کنـد. از سـوی دیگـر، بـا تمرکـز بـر بافتـار توانمنـد، امیـد بـه افزایـش

دقت بازیابی و در نتیجه، افزایش اثربخشی آن بیشتر می‌شود. با این حال، با توجه به محدودبودن مجموعه مدارک آزمایشی مورد بررسی در این پژوهش توصیه می‌شود که این تحقیق بر مجموعه‌ای انبوه از مدارک تکرار شود و نتایج آن با تحقیق حاضر مقایسه گردد.

قدردانی

نویسندگان بر خود لازم می‌دانند از جناب آقای دکتر فخراحمد، استادیار محترم بخش علوم کامپیوتر دانشگاه شیراز به جهت آرای ارزشمند ایشان در طول انجام تحقیق سپاسگزاری نمایند.

فهرست منابع

امیدزاده، راضیه، عبدالمجید موسوی، و حسن نادری. ۱۳۹۲. *افزایش دقت الگوریتم درخت تصمیم در رفع ابهام معنایی کلمات با استفاده از روش ad tree*. اولین همایش منطقه‌ای بهینه‌سازی و روش‌های محاسبهٔ نرم در مهندسی برق و کامپیوتر، صفاشهر، دانشگاه آزاد اسلامی واحد صفاشهر، http://www. civilica. com/Paper-ELECOM01-ELECOM01_080. html

اینگورسن، پیتر. ۱۹۹۲. *تعامل بازیابی اطلاعات*. ترجمهٔ هاجر ستوده. ۱۳۸۹. ویراستار بهاره پورحسن. تهران: کتابدار.

بزم‌آرا، محمد، شهرام جعفری، و علی بزم‌آرا. ۱۳۹۲. *رفع ابهام معنایی در ترجمهٔ ماشینی با استفاده از الگوریتم‌های یادگیری با نظارت*. اولین همایش ملی کاربرد سیستم‌های هوشمند (محاسبات نرم) در علوم و صنایع. قوچان، دانشگاه آزاد اسلامی واحد قوچان، http://www. civilica. com/Paper-AISST01-AISST01_079. html

بلیکی، نورمن. (۱۳۸۴). *طراحی پژوهش‌های اجتماعی*. ترجمهٔ حسن چاوشیان. تهران: نشر نی.

بی‌جن‌خان، محمود، و شهروز مرادزاده. ۱۳۸۹. *هم‌نگاره‌های خط فارسی*. ارائه‌شده در اولین کارگاه پژوهشی زبان فارسی و رایانه. تهران: سمت.

پائو، میراندالی. ۱۹۸۹. *مفاهیم بازیابی اطلاعات*. ترجمهٔ اسدالله آزاد و رحمت‌الله فتاحی. ۱۳۷۹. مشهد: دانشگاه فردوسی مشهد.

چامسکی، نوام. ۲۰۰۰. *زبان و ذهن*. ترجمهٔ کورش صفری. ۱۳۸۷. تهران: هرمس.

حبیب‌پور گتابی، کرم، و رضا صفری شالی. ۱۳۸۸. *راهنمای جامع کاربرد SPSS در تحقیقات پیمایشی*. تهران: نشر لویه.

رخشانفر، محمدرضا. ۱۳۷۱. معانی و ساختار زبان چندمعنایی و هم‌آوایی واژه‌ها. *رشد آموزش زبان* ۳۴:

۶۵-۲۹.

زمانی، محمدعلی. ۱۳۸۷. معرفی و نقد و بررسی کتاب: معناشناسی و بازیابی اطلاعات، هفت گفتار. *مطالعات ملی کتابداری و سازماندهی اطلاعات* ۷۶: ۳۰۷-۳۱۴.

ساراسویک، تفکو. ۲۰۱۱. *ربط در علم اطلاع‌رسانی*. ترجمهٔ حیدر مختاری و عباس میرزایی. ۱۳۸۹. ویراستار اعظم صنعت‌جو. تهران: چاپار.

علوی‌مقدم، بهنام. ۱۳۸۹. روابط واژگانی. *فصلنامه مطالعات برنامه درسی* ۱۴ (۲): ۲۸-۳۱.

فلاحتی فومنی، محمدرضا. ۱۳۸۵. ابهام در ماشین ترجمه. *کتابداری و اطلاع‌رسانی* ۹ (۳): ۲۱-۳۹.

لنکستر، فردریک. ۲۰۰۴. *نمایه‌سازی و چکیده‌نویسی، مبانی نظری و عملی*. ترجمهٔ عباس گیلوری. ۱۳۸۲. تهران: نشر چاپار.

مسعودی، بابک، سعید راحتی قوچانی، و اعظم استاجی. ۱۳۸۹الف. *یک روش بینی برای رفع ابهام معنایی کلمات در زبان فارسی با تأکید بر ویژگی‌های محلی کلمه*. اولین کنفرانس ملی محاسبات نرم و فناوری اطلاعات، ماهشهر، دانشگاه آزاد اسلامی واحد ماهشهر، http://www. civilica. com/Paper-NCSCIT01-NCSCIT01_055. html

_____. ۱۳۸۹ب. *یک مدل بیشینه بی‌نظمی جهت رفع ابهام معنایی کلمات فارسی به‌کمک ویژگی‌های مدل‌سازی موضوع*. شانزدهمین کنفرانس بین‌المللی سالانه انجمن کامپیوتر ایران. تهران، انجمن کامپیوتر، http://www. civilica. com/Paper-CSICC16-CSICC16_074. html

معروفی، افسانه، و عبدالحمید پیله‌ور. ۱۳۹۰. *رفع ابهام از معنی کلمات مبهم فارسی با استفاده از روش‌ها مبتنی بر پیکره و قاموس*. اولین همایش منطقه‌ای رویکردهای نوین در مهندسی کامپیوتر و فناوری اطلاعات. رودسر، دانشگاه آزاد اسلامی واحد رودسر و آمل، http://www. civilica. com/Paper-ROUDSARIT01-ROUDSARIT01_257. html

معناشناسی و بازیابی اطلاعات، هفت گفتار (). ترجمه و تألیف جعفر مهراد، محمدرضا فلاحتی فومنی. ۱۳۸۴. مشهد: کتابخانه رایانه‌ای؛ شیراز: کتابخانه منطقه‌ای علوم و تکنولوژی.

مهراد، جعفر و محمدرضا فلاحتی فومنی (۱۳۸۴). معناشناسی و بازیابی اطلاعات، هفت گفتار. مشهد: کتابخانه رایانه ای؛ شیراز: کتابخانه منطقه‌ای علوم و تکنولوژی.

مینایی بیدگلی، بهروز، احمد اکبری، و مهدی محسنی. ۱۳۸۶. *به‌کارگیری متن‌کاوی در ابهام‌زدایی از هم‌نویسه‌های غیرتکیه‌ای در زبان فارسی*. اولین کنفرانس داده‌کاوی ایران، تهران، دانشگاه صنعتی امیرکبیر، مؤسسه پژوهشی داده‌پردازان گیتا، http://www. civilica. com/Paper-IDMC01-IDMC01_083. html

یوسفی، احمد. ۱۳۷۶. ریزش کاذب در ذخیره و بازیابی اطلاعات. *کتابداری و اطلاع‌رسانی* ۱۳ (۱): ۱-۹.

Abu-Jbara, A., Ezra, J., & Radev, D. 2013. Purpose and polarity of citation: Towards NLP-based bibliometrics. In Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies (pp. 596-606).

Choi, S. H. 2010. Document Clustering Using Reference Titles. 27 (2): 241-252.

Church, K. W., & P. Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics* 16 (1): 22-29.

Fillmore, C. J. 1985. Frames and the semantics of understanding. *Quaderni di semantica* 6 (2): 222-254.

Han, H., L. Giles, H. Zha, C. Li, & K. Tsioutsiouliklis. 2004. Two supervised learning approaches for name disambiguation in author citations. In *Digital Libraries. Proceedings of the 2004 Joint ACM/ IEEE Conference on (pp. 296-305). IEEE. JCDL'04, June 7–11, 2004, Tucson, Arizona, USA*

Harmandas, V., M. Sanderson, & M. D. Dunlop. 1997. Image retrieval by hypertext links. In: *Proceedings of the 20th Annual International ACM SIGIR conference on Research and development in Information Retrieval*. ACM, pp. 295-303. ISBN 0-89791-836-3. SIGIR 97 Philadelphia USA.

Hearst, M. A. 1991. Noun homograph disambiguation using local context in large text corpora. In The Proceedings of the 7th Annual Conference of the UW Centre for the New OED and Text Research: Using Corpora: 185-188. Oxford, UK.

Hindle, D. 1990. Noun classification from predicate-argument structures. Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics: 268-275.

Hjørland, B. 2008. What is knowledge organization (KO)? *Knowledge organization. International journal devoted to concept theory, classification, indexing and knowledge representation* 35 (2/3): 90-100.

Hurford, J. R., B. Heasley, & M. B. Smith. 2007. *Semantics: a coursebook*. New York: Cambridge University Press.

Jeong, Y. K., M. Song, & Y. Ding. 2014. Content-based author co-citation analysis. *Journal of Informetrics* 8 (1): 197-211.

Lee, Y. K., H. T. Ng, & T. K. Chia. 2004. Supervised word sense disambiguation with support vector machines and multiple knowledge sources. In *Senseval-3: third international workshop on the evaluation of systems for the semantic analysis of text (pp. 137-140). Barcelona, Spain.*

Machlup, F. 1983. "Semantic Quirks in Studies of Information," in F. Machlup and U. Mansfield (Eds.), The Study of Information, Pp 641-671. New York: John Wiley.

Makki, R., & M. M. Homayounpour, 2008. Word sense disambiguation of Farsi homographs using thesaurus and corpus. In Advances in Natural Language Processing (pp. 315-323). Springer, Berlin, Heidelberg.

Manning, C. D., P. Raghavan, & H. Schütze. 2008. *Introduction to information retrieval* (Vol. 1, p. 496). Cambridge: Cambridge university press.

Nakov, P. I., A. S. Schwartz, & M. Hearst. 2004. Citances: Citation sentences for semantic analysis of bioscience *text. In Proceedings of the SIGIR'04 workshop on Search and Discovery in Bioinformatics (pp. 81-88). Sheffield, UK.*

Nanba, H., & M. Okumura. 1999. Towards multi-paper summarization using reference information *IJCAI* 99: 926-931).

Qazvinian, V., & D. R. Radev. 2010. Identifying non-explicit citing sentences for citation-based summarization. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 555-564).

Palmer, F. R. 1976. Semantics: a new outline. London: Cambridge University Press.

Qazvinian, V., & D. R. Radev. 2010. Identifying non-explicit citing sentences for citation-based summarization. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 555-564). Uppsala, Sweden

Rezapour, A. R., S. M. Fakhrahmad, & M. H. Sadreddini. 2011. Applying weighted KNN to word sense disambiguation. In *Proceedings of the World Congress on Engineering* (Vol. 3, pp. 6-8), UK.

Riahi, N., & F. Sedghi, 2012. A semi-supervised method for Persian homograph disambiguation.

In Electrical Engineering (ICEE), 2012 20th Iranian Conference on (pp. 748-751). IEEE.

Searle, J. R. 1984. Intentionality and its place in nature. *Dialectica* 38 (2-3): 87-99.

Small, H. 2011. Interpreting maps of science using citation context sentiments: a preliminary investigation. *Scientometrics* 87 (2): 373-388.

Tang, J., A. C. M. Fong, B. Wang, & J. Zhang. 2012. A unified probabilistic framework for name disambiguation in digital library. *Knowledge and Data Engineering, IEEE Transactions* 24 (6): 975-987.

Tesprasit, V., P. Charoenpornsawat, & V. Sornlertlamvanich. 2003. A context-sensitive homograph disambiguation in Thai text-to-speech synthesis. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003--short papers-Volume 2 (pp. 103-105). Association for Computational Linguistics.*

Tong, T., D. Dinakarpandian, & Y. Lee. 2009. Literature clustering using citation semantics. In *System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on System Sciences (pp*. 1-10). IEEE.

Tsai, C. T., G. Kundu, & D. Roth. 2013. Concept-based analysis of scientific literature. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management* (pp. 1733-1738). *New York, NY, USA: ACM.*

Twilley, L. C., P. Dixon, D. Taylor, & K. Clark. 1994. University of Alberta norms of relative meaning frequency for 566 homographs. *Memory & Cognition* 22 (1): 111-126.

W Weizenbaum, J. 1976. *Computer power and human reason*. San Francisco: W. H. Freeman.

هاجر ستوده

دارای مدرک دکتری علم اطلاعات و دانش‌شناسی است. ایشان هم‌اکنون دانشیار دانشگاه شیراز است.

علم‌سنجی، سازماندهی و مدیریت دانش، و بازیابی اطلاعات از جمله علایق پژوهشی وی است.

مژگان هوشیار

دارای مدرک کارشناسی ارشد علم اطلاعات و دانش‌شناسی از دانشگاه شیراز است.

بازیابی اطلاعات و علم‌سنجی از جمله علایق پژوهشی وی است.