



وزارت علوم، تحقیقات، و فناوری
پژوهشگاه علوم و فناوری اطلاعات ایران
(ایرانداک)

طرح پژوهشی

**توسعه رویکرد خوشه‌بندی برای پیشنهاد اسناد علمی فارسی
بر مبنای داده‌های فارسی پایگاه اطلاعات علمی ایران (گنج)**

مجری

مرضیه زرین بال ماسوله

همکار

آزاده محبی

بهمن ۱۴۰۰

بنام خداوند جان و

حقوق: پژوهشگاه علوم و فناوری اطلاعات ایران

طرح پژوهشی «توسعه رویکرد خوشه‌بندی برای پیشنهاد اسناد علمی فارسی بر مبنای داده‌های فارسی پایگاه اطلاعات علمی ایران (گنج)» پیرو قرارداد شماره ۳۳/۱۲۶۱۸ مورخ ۱۳۹۸/۱۱/۰۷ میان پژوهشگاه علوم و فناوری اطلاعات ایران (کارفرما) و خانم مرضیه زرین بال ماسوله (مجری) و خانم آزاده محبی (همکار) اجرا شده است.

این گزارش و تمامی حقوق مادی آن بر اساس «قانون حمایت حقوق مؤلفان و مصنفان و هنرمندان، مصوب سال ۱۳۸۴ و اصلاحیه‌های بعدی آن و همچنین آیین‌نامه‌های اجرایی این قانون متعلق به پژوهشگاه علوم و فناوری اطلاعات ایران است و هر گونه استفاده از تمامی یا پاره‌ای از آن، شامل: نقل قول، تکثیر، انتشار، کاربرد نتایج، تکمیل و مانند آنها به صورت چاپی، الکترونیکی یا وسایل دیگر فقط با اجازه کتبی پژوهشگاه امکانپذیر است. نقل قول در حد هزار واژه در انتشارات علمی مانند کتاب و مدرک با درج اطلاعات کامل کتابشناختی، نیازی به مجوز پژوهشگاه ندارد.

صحت مندرجات گزارش بر عهده مجری طرح پژوهشی است.



وزارت علوم، تحقیقات و فناوری
پژوهشگاه علوم و فناوری اطلاعات ایران
ایرانداک

شماره:

۳۳/۱۰۴۰۶

تاریخ: ۱۴۰۰/۱۲/۰۹

پژوهشی - مدیریت دانش - آموزشی

ندارد

به نام خدا

ایرانداک در نیم‌قرن دوم کار خود همچنان به‌طور و کوشا، همراه هر پژوهش، پژوهشگر، و سیاست‌گذار

۱۳۴۷ - ۱۴۰۰

گواهی

انجام طرح پژوهشی:

- ✧ عنوان: توسعه رویکرد خوشه‌بندی برای پیشنهاد اسناد علمی فارسی بر مبنای داده‌های فارسی پایگاه اطلاعات علمی ایران (گنج)
- ✧ مجری: دکتر مرضیه زرین‌بال ماسوله
- ✧ همکار: دکتر آزاده مجبی
- ✧ شماره و تاریخ قرارداد: ۳۳/۱۲۶۱۸ (۱۳۹۸/۱۱/۷)
- ✧ تاریخ شروع: ۱۳۹۸/۱۱/۷
- ✧ تاریخ پایان (تصویب گزارش پایانی در شورای پژوهش): ۱۴۰۰/۱۲/۳ (نشست ۴۴۷)
- ✧ ناظر: دکتر مرتضی صابری

در پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) گواهی می‌شود. از سوی ایرانداک، از مجری، همکار، و ناظر گرامی این طرح ارزشمند سپاسگزارم.

با آرزوی پیروزی
پروژه شهریاری
معاون پژوهش و آموزش

توسعه رویکرد خوشه‌بندی برای پیشنهاد اسناد علمی فارسی بر مبنای داده‌های فارسی
پایگاه اطلاعات علمی ایران (گنج)

مدیر طرح پژوهشی: مرضیه زرین بال ماسوله، پژوهشگاه علوم و فناوری اطلاعات ایران

همکار طرح پژوهشی: آزاده محبی، پژوهشگاه علوم و فناوری اطلاعات ایران

نشانی: تهران، خیابان انقلاب، چهارراه فلسطین، شماره ۱۰۹۰،

صندوق پستی: ۱۳۱۵۷۷۳۳۱۴، تلفن: ۰۲۱۶۶۴۹۴۹۸۰، دورنگار: ۰۲۱۶۶۴۹۴۹۸۰

وب‌گاه: www.irandoc.ac.ir

رایانامه: zarinbal@irandoc.ac.ir، mohebi@irandoc.ac.ir

سیستم‌های پیشنهاددهنده سیستم‌هایی هستند که با ارائه پیشنهاد‌های متناسب با نیاز کاربر به او در تصمیم‌گیری کمک می‌کنند. نیازهای کاربر با کمک روش‌های مختلفی نظیر بهره‌گیری از عبارت پرس‌وجوی کاربر، تحلیل سند منتخب کاربر و یا تحلیل رفتار کاربر و یا کاربران مشابه استخراج می‌شود. سیستم پیشنهاددهنده مبتنی بر دانش یا مبتنی بر محتوی یکی از انواع سیستم‌های پیشنهاددهنده است و برای توسعه آن خصوصاً در سال‌های اخیر از رویکردهای نوین نظیر یادگیری ماشین و تشخیص الگو استفاده می‌شود. خوشه‌بندی متن یکی از روش‌های بسیار پرکاربرد در یادگیری ماشین است که در آن اسناد مشابه براساس ویژگی‌های مناسب و شباهت بین اسناد خوشه‌بندی می‌شوند. لذا استخراج و انتخاب این ویژگی‌ها برای اسناد و خوشه‌بندی اسناد از جمله چالش‌های اصلی محسوب می‌شوند. همچنین، اغلب سیستم‌های پیشنهاددهنده متن و الگوریتم‌های خوشه‌بندی برای زبان انگلیسی توسعه پیدا کرده‌اند و به راحتی برای سایر زبان‌ها، مانند زبان فارسی، قابل بکارگیری نیستند. بنابراین هدف اصلی در این طرح، توسعه و بکارگیری رویکرد خوشه‌بندی برای پیشنهاد اسناد علمی فارسی به کاربر است. برای ارزیابی رویکرد توسعه داده شده از فراداده‌های پایان‌نامه‌ها و رساله‌های فارسی پایگاه گنج استفاده شد. براساس ادبیات مربوط به خوشه‌بندی داده‌های بزرگ، سه معیار شباهت مینکویسکی، کسینوسی و آنتروپی نسبی به عنوان معیارهای فاصله یا شباهت انتخاب و با کمک این معیارها سه الگوریتم خوشه‌بندی K-means، FCM و REFCM در نرم‌افزار Python پیاده‌سازی شدند. براساس نتایج بدست آمده، الگوریتم FCM (فاصله کسینوسی) و REFCM نتایج بهتری نسبت به الگوریتم K-means داشتند.

کلید واژه‌ها: اسناد علمی فارسی؛ یادگیری ماشین؛ خوشه‌بندی عمیق؛ آنتروپی نسبی

عنوان	شماره صفحه
چکیده.....	أ
فهرست مطالب.....	ب
فهرست جدول‌ها.....	ث
فهرست شکل‌ها.....	ج
فصل اول: بیان مسئله.....	۲
۱-۱ بیان مسئله.....	۲
۱-۲ هدف‌ها.....	۵
۱-۳ پرسش‌ها.....	۵
۱-۴ روش پژوهش.....	۵
فصل دوم: خوشه‌بندی.....	۹
۲-۱ مقدمه.....	۹
۲-۲ سیستم‌های پیشنهاددهنده.....	۹
۲-۳ متن کاوی.....	۱۳
۲-۳-۱ داده‌های متنی.....	۱۵
۲-۳-۲ نمایش داده‌های متنی.....	۱۸
۲-۴ خوشه‌بندی.....	۱۹
۲-۵ خوشه‌بندی اسناد.....	۲۱
۲-۵-۱ گام ۱: بازنمایی مسئله.....	۲۱
۲-۵-۲ گام ۲: معیار شباهت.....	۲۷
۲-۵-۳ گام ۳: خوشه‌بندی داده‌ها.....	۲۹
۲-۵-۴ گام ۴: خلاصه‌سازی نتایج و داده‌ها.....	۳۴
۲-۵-۵ گام ۵: ارزیابی.....	۳۴

۳۵	۶-۲ خوشه‌بندی اسناد کوتاه
۳۷	فصل سوم: رویکرد خوشه‌بندی پیشنهادی هوش مصنوعی
۳۷	۱-۳ مقدمه
۳۷	۲-۳ داده‌های مورد پردازش
۳۹	۳-۳ رویکرد خوشه‌بندی پیشنهادی
۳۹	۱-۳-۳ گام ۱: بازنمایی مسئله
۴۰	۲-۳-۳ گام ۲: معیار شباهت
۴۱	۳-۳-۳ گام ۳: خوشه‌بندی داده‌ها
۴۳	۴-۳-۳ گام ۴: ارزیابی
۴۵	فصل چهارم: پیاده‌سازی رویکرد خوشه‌بندی پیشنهادی هوش مصنوعی
۴۵	۱-۴ مقدمه
۴۵	۲-۴ گام ۱: بازنمایی مسئله
۴۶	۳-۴ گام ۲: معیار شباهت
۴۶	۴-۴ گام ۳: خوشه‌بندی داده‌ها
۴۷	۵-۴ گام ۴: ارزیابی
۴۹	نتیجه‌گیری
۵۰	پیشنهاد پژوهشی و کاربردی
۵۲	منابع
Abstract	57

جدول ۳-۱- تعداد اسناد در هر حوزه موضوعی.....	۳۸
جدول ۳-۲- شبه کد الگوریتم k-means.....	۴۱
جدول ۴-۱- نتایج بدست آمده برای $c=13$	۴۷
جدول ۴-۲- نتایج بدست آمده برای $c=11$	۴۸
جدول ۴-۳- تفسیر خوشه‌های بدست آمده در حالت $c=11$	۴۸

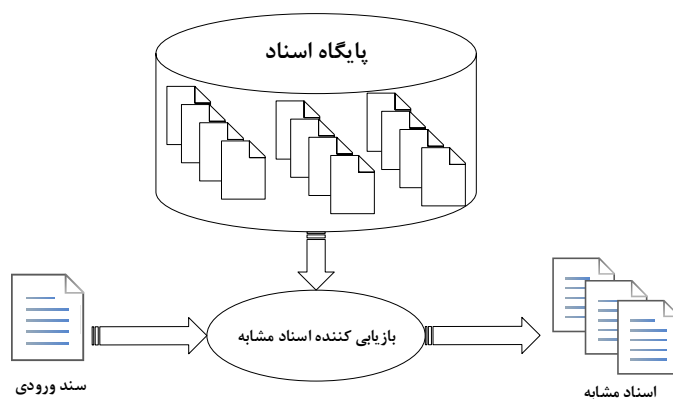
شکل ۱-۱- پیشنهاد اسناد مشابه.....	۲
شکل ۲-۱- سیستم پیشنهاد اسناد مشابه در اسکوپوس.....	۳
شکل ۳-۱- فرآیند کلی جستجو در آپاچی سولر (زرین بال، ۱۳۹۷).....	۴
شکل ۴-۱- ماتریس آشفتگی.....	۶
شکل ۱-۲- سیستم‌های پیشنهاددهنده و منبع دانش بکاررفته برای ساخت (Burke, 2007).....	۱۱
شکل ۲-۲- ارتباط زمینه‌های متن کاوی (FarsiYar, 2020).....	۱۵
شکل ۳-۲- نمونه‌ای از پردازش‌ها در درک زبان طبیعی (Zhai & Massung, 2016).....	۱۷
شکل ۴-۲- طبقه‌بندی الگوریتم‌های خوشه‌بندی (Saxena et al., 2017).....	۱۹
شکل ۵-۲- خوشه‌بندی تجمیعی و خوشه‌بندی تقسیمی (Saxena et al., 2017).....	۲۰
شکل ۶-۲- ساختار دو مدل CBOW و Skip-gram (محبی، جلالی‌منش، ۱۳۹۷).....	۲۴
شکل ۷-۲- چارچوب روش خوشه‌بندی عمیق (Aljalbout et al., 2019).....	۳۲
شکل ۸-۲- رویکردهای خوشه‌بندی عمیق بر اساس پژوهش (Aljalbout et al., 2019).....	۳۳
شکل ۱-۳- برشی از مجموعه داده.....	۳۸
شکل ۱-۴- بخشی از بازنمایی داده‌های ورودی با کمک روش SBERT.....	۴۶

فصل اول

بیان مسئلہ

۱-۱ بیان مسئله

سیستم‌های پیشنهاددهنده سیستم‌هایی هستند که با هدف ارائه پیشنهادهایی به کاربر توسعه یافته‌اند، این پیشنهادها متناسب با نیاز کاربر بوده و به او در تصمیم‌گیری برای انتخاب موارد مورد نیاز یا مورد علاقه کمک می‌کنند. نیازها یا علائق کاربر با کمک روش‌های مختلفی استخراج می‌شود، نظیر بهره‌گیری از عبارت پرس‌وجوی کاربر و تحلیل سند منتخب کاربر و استفاده از ویژگی‌های خاص آن برای پیشنهاد اسناد مشابه با سند (اسناد) منتخب کاربر (شکل ۱-۱).



شکل ۱-۱- پیشنهاد اسناد مشابه

توسعه رویکرد خوشه‌بندی برای پیشنهاد اسناد علمی فارسی ... □ ۳

در کنار سامانه‌های تجاری نظیر Amazon.com, YouTube, Netflix, Yahoo, Tripadvisor و IMDb، برخی از سامانه‌های جستجوی اسناد علمی نظیر اسکوپوس، امرالد و پاب‌مد از سیستم‌های پیشنهاددهنده برای بهبود تجربه کاربران خود و تعامل بهتر با آنان استفاده می‌کنند. نمونه‌ای چنین سیستمی در پایگاه اسکوپوس در شکل ۲-۱ نشان داده شده است.

The screenshot shows the article page for "Robust credibilistic fuzzy local information clustering with spatial information constraints" in the journal "Digital Signal Processing". The authors are Chengmao Wu and Xiaoqiang Yang. The page includes a sidebar with recommended articles and a list of citing articles. The abstract is partially visible at the bottom.

Download PDF Share Export Search ScienceDirect Advanced

Digital Signal Processing
Volume 97, February 2020, 102615

Robust credibilistic fuzzy local information clustering with spatial information constraints

Chengmao Wu, Xiaoqiang Yang

Show more

<https://doi.org/10.1016/j.dsp.2019.102615> Get rights and content

Abstract

Credibilistic fuzzy clustering is a novel data analysis method. Aiming at the shortcoming that credibilistic fuzzy clustering algorithm (CFCM) lacks the

Recommended articles

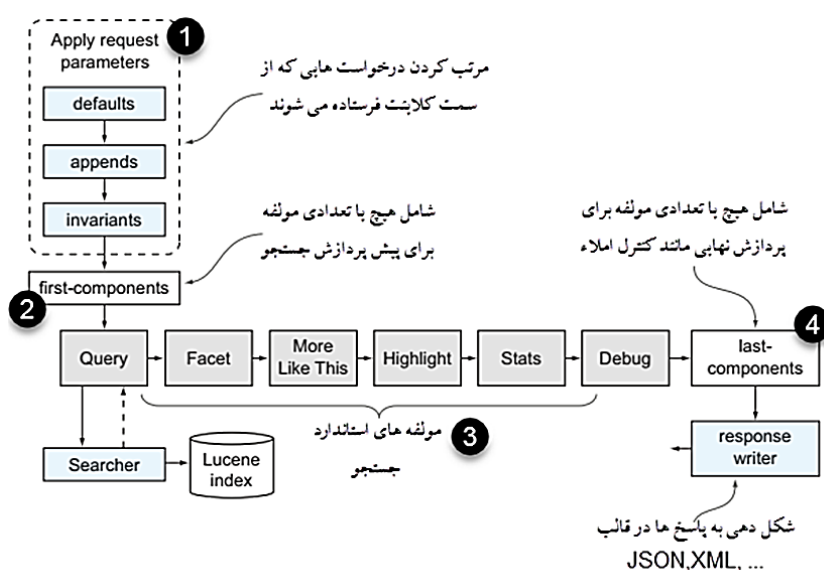
- Adaptive transmit array sidelobe control usi...
Digital Signal Processing, Volume 97, 2020, Article ...
Download PDF View details
- Use of PALM for ℓ_1 sparse matrix factorizati...
Digital Signal Processing, Volume 97, 2020, Article ...
Download PDF View details
- DOA estimation via ULA with mutual coupli...
Digital Signal Processing, Volume 97, 2020, Article ...
Download PDF View details

1 2 Next >

Citing articles (0)

شکل ۲-۱- سیستم پیشنهاد اسناد مشابه در اسکوپوس

در کتابخانه آپاچی سولر (هسته اصلی سامانه گنج) نیز چنین قابلیت وجود دارد که توسط مولفه More like this و تنظیمات مورد نیاز در Schema فعال می‌شود (شکل ۳-۱) با این حال، این امر در حال حاضر برای سامانه گنج فعال نیست (زرین بال، ۱۳۹۷). البته این مولفه در سولر براساس عبارت پرس‌وجوی کاربر تعریف شده است. به بیانی دیگر، سولر اسناد مشابه با سند منتخب کاربر را براساس عبارت پرس‌وجوی او جستجو کرده و از ویژگی‌های سیستم‌های پیشنهاددهنده استفاده نمی‌کند (Grainger and Potter 2014).



شکل ۳-۱- فرآیند کلی جستجو در آپاچی سولر (زرین بال، ۱۳۹۷)

به بیانی دیگر، این مولفه سولر محدود به واژگان موجود در عبارت پرس‌وجوی کاربر بوده و در صورتیکه این عبارت دارای واژگان کلی یا کمی باشد، اسناد مناسبی پیشنهاد نخواهد شد. یک رویکرد برای بهبود این موضوع، استفاده از سیستم‌های پیشنهاددهنده است. سیستم‌های پیشنهاددهنده انواع مختلف دارند که از بین آنها سیستم‌های پیشنهاددهنده مبتنی بر محتوا می‌تواند برای پیشنهاد اسناد مشابه بکار گرفته شوند. یکی از رویکردهای پرکاربرد برای بهره‌گیری از این نوع سیستم پیشنهاددهنده، رویکرد خوشه‌بندی^۱ اسناد مشابه و یافتن مشابه‌ترین سند با سند انتخابی کاربر است. لیکن استخراج و انتخاب ویژگی‌های مناسب برای اسناد و یافتن شباهت بین آنها نیازمند بکارگیری روش‌های متن‌کاوی و یادگیری ماشین^۲ برای متون بوده و از جمله چالش‌های اصلی است. علاوه بر آن، اغلب این سیستم‌ها برای زبان انگلیسی توسعه پیدا کرده‌اند و به راحتی برای سایر زبان‌ها قابل بکارگیری نیستند.

به دلیل جذابیت اضافه شدن چنین سیستمی به سامانه گنج، انجام طرح پژوهشی با عنوان «توسعه رویکرد خوشه‌بندی برای پیشنهاد اسناد علمی فارسی بر مبنای داده‌های فارسی» پیشنهاد می‌شود که برای انجام آن از داده‌های مربوط به پایان‌نامه‌ها و رساله‌های (پارساهای) فارسی موجود در سامانه گنج در حوزه تخصصی فنی و مهندسی استفاده می‌شود. هدف از انجام آن نیز پیشنهاد اسناد علمی فارسی

^۱ Clustering

^۲ Machine learning

بر اساس رویکرد خوشه‌بندی است که در طی آن پارساها بر اساس ویژگی‌های شناخته شده و روش انتخاب شده، خوشه‌بندی می‌شوند. چنین سیستمی (با اندکی تغییر) کاربردهای فراوانی در سامانه‌های مختلف خواهد داشت که از آن میان می‌توان به کاربرد آن در سامانه پیشینه پژوهش و پژوهشنامه پردازش و مدیریت اطلاعات اشاره نمود.

۲-۱ هدف‌ها

بر اساس موارد مطرح شده، اهداف اصلی طرح پژوهشی مذکور بشرح زیر است:

□ توسعه رویکرد خوشه‌بندی برای پیشنهاد اسناد علمی فارسی

جهت دستیابی به اهداف فوق اهداف فرعی طرح پژوهشی به شرح زیر تعیین می‌گردند:

□ توسعه رویکردی برای خوشه‌بندی پارساهای موجود در پایگاه گنج

□ انتخاب ویژگی‌های مناسب اسناد با هدف خوشه‌بندی

□ انتخاب معیار شباهت مناسب برای خوشه‌بندی و پیشنهاد اسناد مشابه با سند انتخابی کاربر.

۳-۱ پرسش‌ها

جهت دستیابی به اهداف فوق در این طرح به دنبال دستیابی به پاسخ پرسش‌های زیر هستیم:

□ چه رویکردهای مطرحی برای خوشه‌بندی اسناد وجود دارند؟ این رویکردها با چه هدفی

توسعه داده شده‌اند و دارای چه محدودیت‌هایی هستند؟ از چه رویکردهای دیگری می‌توان

استفاده کرد و یا چگونه می‌توان آنها را توسعه داد؟

□ از چه ویژگی‌هایی برای خوشه‌بندی اسناد استفاده می‌شود؟ از کدامیک می‌توان برای

پارساها استفاده کرد؟

□ چه معیارهایی برای محاسبه شباهت بین اسناد وجود دارند؟ از کدامیک می‌توان برای پارساها

استفاده کرد؟

۴-۱ روش پژوهش

در پژوهش حاضر ابتدا ویژگی‌های مورد نیاز برای خوشه‌بندی اسناد و معیار شباهت از ادبیات استخراج شده و با نداشت آنها بر اساس فارسی، رویکرد خوشه‌بندی پیشنهادی ارائه می‌شود. اعتبار چنین سیستم‌هایی در دو مرحله کلان سنجیده می‌شود، سنجش اعتبار اجزای سیستم به صورت مجزا و

سنجش اعتبار سیستم به صورت کل و یکپارچه. برای سنجش اعتبار نتایج خوشه‌بندی معیارهای مشخصی وجود دارد: دقت^۱ (رابطه (۱-۱))، صحت^۲ (رابطه (۲-۱))، حساسیت^۳ (رابطه (۳-۱))، تشخیص‌پذیری^۴ (رابطه (۴-۱)) و ماتریس آشفتگی^۵ (شکل ۴-۱). برای محاسبه این معیارها داده‌ها به دو بخش تقسیم می‌شوند، بخشی برای آموزش (حدود ۷۰٪ داده‌ها) و بخشی برای آزمایش (حدود ۳۰٪ داده‌ها).

$$accuracy = \frac{\text{No. of true positives} + \text{No. of true negatives}}{\text{No. of true positives} + \text{false positives} + \text{false negatives} + \text{true negatives}} \quad (1-1)$$

$$precision = \frac{\text{No. of true positives}}{\text{No. of true positives} + \text{false positives}} \quad (2-1)$$

$$Sensitivity = \frac{\text{No. of true positives}}{\text{No. of true positives} + \text{false negatives}} \quad (3-1)$$

$$specificity = \frac{\text{No. of true negatives}}{\text{No. of false positives} + \text{true negatives}} \quad (4-1)$$

		شرط مثبت	شرط منفی	
خروجی	خروجی مثبت	مثبت درست (True Positive)	مثبت اشتباه (False positive)	پیش‌بینی مثبت (دقت)
	خروجی منفی	منفی غلط	منفی درست	پیش‌بینی منفی
		حساسیت	تشخیص‌پذیری	صحت

شکل ۴-۱- ماتریس آشفتگی

^۱ Precision

^۲ Accuracy

^۳ Sensitivity

^۴ Specificity

^۵ Confusion matrix

برای سنجش اعتبار کل سیستم نیز می‌توان از معیارهایی چون صحت، پوشش^۱ (رابطه ۵) و تنوع^۲ در ارائه پیشنهادها به کاربر استفاده نمود (Ricci et al. 2011).

(۵-۱)

$$H = - \sum_{i=1}^n p(i) \log(p(i))$$

بر این اساس، پژوهش حاضر در قالب دو مرحله اصلی زیر صورت خواهد پذیرفت:

مرحله اول: مطالعه ادبیات رویکردهای خوشه‌بندی متن در سیستم‌های پیشنهاددهنده

- مطالعه رویکردهای مطرح خوشه‌بندی متن و ویژگی‌های هر روش
- بررسی ویژگی‌های مورد استفاده در رویکردهای خوشه‌بندی متن برای سیستم‌های پیشنهاددهنده
- بررسی معیارهای شباهت مورد استفاده در رویکردهای خوشه‌بندی متن برای سیستم‌های پیشنهاددهنده

مرحله دوم: پیاده‌سازی و ارزیابی

- گردآوری داده‌های مورد نیاز
- پیشنهاد رویکرد خوشه‌بندی مناسب برای بکارگیری در سیستم پیشنهاددهنده متنی
- نگاشت ویژگی‌ها و معیارهای شباهت بر داده‌های موجود
- پیاده‌سازی رویکرد خوشه‌بندی پیشنهادی به صورت آزمایشی و ارزیابی

^۱ Coverage

^۲ Diversity

فصل دوم

خوشه‌بندی

۱-۲ مقدمه

این فصل به مطالعه رویکردهای مطرح خوشه‌بندی متن اختصاص دارد. بدین منظور ابتدا توضیح مختصری در رابطه با چرایی نیاز به خوشه‌بندی متن از منظر سیستم‌های پیشنهاددهنده پرداخته شده و سپس اصول کلی متن‌کاوی تشریح می‌شود. در ادامه ابتدا روش‌های کلان خوشه‌بندی اشاره شده و سپس روش‌ها و گام‌های مورد استفاده در خوشه‌بندی متن بیان می‌شود.

۲-۲ سیستم‌های پیشنهاددهنده

همانطور که اشاره شد، سیستم‌های پیشنهاددهنده سیستم‌هایی هستند که با هدف ارائه پیشنهاد به کاربر توسعه یافته‌اند. این پیشنهادها متناسب با نیاز کاربر بوده و به او در تصمیم‌گیری برای انتخاب موارد مورد نیاز یا مورد علاقه کمک می‌کنند. به بیانی دیگر، این سیستم‌ها فهرستی رتبه‌بندی شده از مواردی را که با احتمال زیاد کاربر به آن‌ها نیاز یا علاقه دارد ارائه می‌کند. این فهرست می‌تواند بر اساس پیشنهاد کاربران مشابه، تحلیل رفتار فعلی کاربر، شناسایی موارد مشابه با مورد انتخاب شده کاربر و یا ترکیبی از این موارد شکل گیرد (Ricci et al. 2011).

از چنین سیستم‌هایی در اغلب سامانه‌های تجاری استفاده شده است که از آن میان می‌توان به Amazon.com, YouTube, Netflix, Yahoo, Tripadvisor و IMDb اشاره کرد. هدف این سامانه‌ها از بکارگیری چنین سیستمی نگهداشت کاربران فعلی و جذب کاربران جدید، افزایش اعتماد

کاربران، افزایش فروش محصولات و یا خدمات، افزایش فروش محصولات و یا خدمات متنوع و درک بهتر نیازهای کاربران است (Ricci et al. 2011).

چهار دسته مهم انواع سیستم‌های پیشنهاددهنده، بر مبنای منبع دانش بکاررفته برای ساخت، عبارتند از (Burke 2007):

- سیستم‌های پیشنهاددهنده همکارانه^۱: در این سیستم‌ها، کاربران مشابه به کاربر فعلی شناسایی شده و پیشنهاد بر اساس انتخاب کاربران مشابه به او ارائه می‌شود.
- سیستم‌های پیشنهاددهنده مبتنی بر محتوی^۲: در این سیستم‌ها، ویژگی‌های کالاها و یا خدمات منتخب کاربر فعلی بررسی شده و مواردی به او پیشنهاد می‌شود که با این ویژگی‌ها همخوانی یا شباهت داشته باشند.
- سیستم‌های پیشنهاددهنده مبتنی بر دانش^۳: در این سیستم‌ها نیازهای کاربر بر اساس رفتار و یا عبارت پرس‌وجوی او استنباط شده و پیشنهادها ارائه می‌شود.
- سیستم‌های پیشنهاددهنده جمعیت‌شناختی^۴: در این سیستم‌ها بر اساس ویژگی‌های جمعیت‌شناختی کاربر پیشنهادها به او ارائه می‌شود.

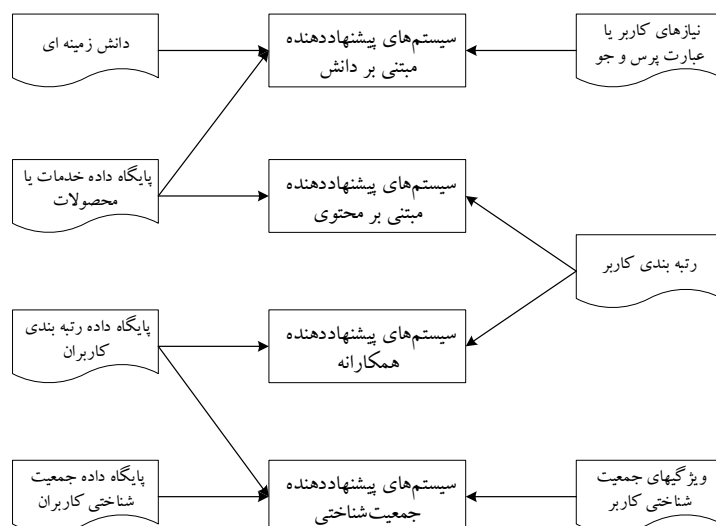
این سیستم‌ها به همراه منبع دانش بکاررفته برای ساخت در شکل ۱-۲ نشان داده شده است.

^۱ Collaborative

^۲ Content-based

^۳ Knowledge-based

^۴ Demographic



شکل ۱-۲- سیستم‌های پیشنهاددهنده و منبع دانش بکاررفته برای ساخت (Burke, 2007)

از بین این سیستم‌ها، سه سیستم پیشنهاددهنده مبتنی بر محتوی، همکارانه و جمعیت‌شناختی از مشکل شروع سرد^۱ رنج می‌برند. به عبارت بهتر، در این سه سیستم باید اطلاعات اولیه از کاربران مشابه و کاربر فعلی جمع‌آوری و تحلیل شده تا بتوان پیشنهادهایی را ارائه داد. در سیستم‌های پیشنهاددهنده مبتنی بر دانش نیز بایستی دانش زمینه‌ای از پایگاه داده و دانش مورد نیاز از کاربر استخراج شده و سپس پیشنهادها ارائه شود (Burke 2007). لذا انتخاب و توسعه هریک از این سیستم‌ها نیازمند بررسی ویژگی‌های مورد نیاز و شرایط موجود است.

در اغلب پایگاه‌های اطلاعاتی مبتنی بر سند، از سیستم‌های پیشنهاددهنده مبتنی بر دانش یا مبتنی بر محتوی استفاده شده و بر اساس مرور ادبیات انجام شده توسط پارک و همکاران رویکردهای بکاررفته در توسعه این سیستم‌ها عبارتند از (Park et al. 2012):

- داده کاوی
- تحلیل لینک^۲
- قواعد انجمنی^۳

^۱ Cold-start

^۲ Link analysis

^۳ Association rules

□ روش‌های نزدیک‌ترین همسایگی^۱

□ شبکه‌های عصبی

□ رگرسیون

□ روش‌های ابتکاری و فراابتکاری^۲.

سیستم‌های متعددی نیز با کمک این رویکردها توسعه داده شده‌اند که از آن میان می‌توان به موارد زیر اشاره نمود:

□ ترکیبی از سیستم پیشنهاددهنده مبتنی بر محتوی و دانش برای پیشنهاد سرفصل‌های درسی دوره‌های آنلاین با کمک روش نزدیک‌ترین همسایگی (Salehi, Pourzaferani, and Razavi 2013)

□ سیستم پیشنهاددهنده مبتنی بر محتوی برای پیشنهاد اسناد شبیه هم براساس تحلیل لینک (De Nart and Tasso 2014)

□ سیستم پیشنهاددهنده مبتنی بر محتوی برای پیشنهاد اسناد در حوزه علوم کامپیوتر با کمک روش‌های نزدیک‌ترین همسایگی و بیزین (Saleh, El Desouky, and Ali 2015)

□ سیستم پیشنهاددهنده مبتنی بر محتوی برای پیشنهاد صفحات وب مشابه براساس مدل‌سازی موضوعی و دسته‌بندی (Sundermann et al. 2016)

□ سیستم پیشنهاددهنده همکارانه متن با کمک روش شبکه‌های عصبی (D. Kim et al. 2017)

□ سیستم پیشنهاددهنده مبتنی بر محتوی برای پیشنهاد مقالات حوزه علوم کامپیوتر با کمک روش رگرسیون (D. Wang et al. 2018)

□ سیستم پیشنهاددهنده مبتنی بر دانش برای پیشنهاد کتاب‌های مشابه با کمک روش قواعد انجمنی (Jomsri 2018)

□ سیستم پیشنهاددهنده متن‌های حاوی کلمات نامناسب با کمک روش شبکه‌های عصبی (Xu et al. 2019)

□ سیستم پیشنهاددهنده مبتنی بر محتوی مقالات مشابه با کمک خوشه‌بندی اسناد (Pradhan and Pal 2019)

^۱ Nearest methods

^۲ Heuristic and metaheuristic methods

□ ترکیبی از سیستم‌های پیشنهاددهنده همکارانه و مبتنی بر محتوی برای پیشنهاد پرسش و پاسخ^۱ مناسب به کاربران با کمک خوشه‌بندی اسناد (M. Li et al. 2020).

در سال‌های اخیر به دلیل ارتقاء سیستم‌های محاسباتی و همچنین به دلیل افزایش حجم داده‌ها در اغلب سیستم‌های پیشنهاددهنده از رویکردهای نوین داده کاوی نظیر یادگیری ماشین و روش‌های تشخیص الگو^۲ استفاده می‌شود. در این روش‌ها هدف اصلی یافتن الگو و شباهت بین داده‌هاست و به دو شیوه کلی بانظارت^۳ و بدون نظارت^۴ طبقه‌بندی می‌شوند. هدف در شیوه بانظارت دسته‌بندی^۵ داده‌ها براساس دانش یا برچسب‌های^۶ موجود در دسته‌های از پیش تعیین شده است درحالی که در شیوه بدون نظارت یا خوشه‌بندی، دانش یا برچسب‌هایی از پیش تعیین شده وجود ندارد و داده‌ها صرفاً براساس شباهت به خوشه‌هایی تعلق می‌گیرند (Zarinbal, Zarandi, and Turksen 2015). در سیستم‌های پیشنهاددهنده توسعه یافته براساس شیوه بانظارت، بایستی دانش یا برچسب‌هایی از پیش وجود داشته باشد. وجود این برچسب‌ها در مورد پایگاه داده‌هایی با حوزه‌های وسیع و متنوع، دشوار و گاهی غیرممکن است. لذا در این سیستم‌ها پیشنهاد می‌شود از روش‌های بدون نظارت استفاده شود. شیوه کلی در این سیستم‌ها آن است که ابتدا ویژگی‌های توصیف کننده سند استخراج شده و سپس با کمک این ویژگی‌ها، اسناد براساس شباهت خوشه‌بندی می‌شوند. در انتها نیز اسناد مشابه در یک خوشه و یا اسناد مکمل در خوشه‌های متفاوت به کاربر پیشنهاد می‌شود. به بیانی دیگر، ابتدا با کمک روش‌های متن کاوی^۷، اسناد تحلیل شده و ویژگی‌های آنها استخراج می‌شود و سپس با کمک روش‌های یادگیری بدون نظارت، اسناد خوشه‌بندی خواهند شد.

در ادامه ابتدا روش‌های متن کاوی و ویژگی‌های اسناد متنی بررسی شده و سپس روش‌های خوشه‌بندی متن تبیین خواهند شد.

۳-۲ متن کاوی

^۱ Question and answer (Q&A)

^۲ Pattern recognition

^۳ Supervised

^۴ Unsupervised

^۵ Classification

^۶ Lable

^۷ Text mining

متن کاوی و یا تحلیل متن فرایند پیدا کردن قوانین و الگوهای غیربديهی، جدید (از قبل نامشخص) و یا مخفی از انبوه مستندات است. به بیانی دیگر، متن کاوی فرایند تحلیل و اکتشاف انبوهی از متون غیر ساخت یافته است با هدف شناسایی مفاهیم، الگوها، موضوعات، کلیدواژه‌ها و دیگر ویژگی‌های داده‌های متنی. متن کاوی دارای ۷ زمینه اصلی است (FarsiYar 2020):

- دسته‌بندی اسناد: فرآیند انتخاب بهترین برچسب برای اسناد بدون برچسب از مجموعه برچسب‌های از قبل مشخص شده است. این کار با استفاده از مدلی انجام می‌شود که از روی متون برچسب‌گذاری شده (داده‌های آموزشی) یاد گرفته و ساخته شده است.
- خوشه‌بندی اسناد: به فرآیند گروه‌بندی مستندات مشابه درون خوشه‌های مختلف خوشه‌بندی می‌گویند.
- بازیابی اطلاعات^۱: فرآیند شاخص‌گذاری^۲، جستجو و بازیابی و استخراج مستندات از بین مجموعه داده‌های عظیم متنی با توجه به عبارت پرس‌وجو^۳ را بازیابی اطلاعات می‌گویند.
- وب کاوی^۴: فرآیند داده و متن کاوی روی محتوای صفحات و ارتباطات بین صفحات وب را وب کاوی می‌گویند.
- استخراج اطلاعات^۵: به فرآیند شناسایی و استخراج موجودیت‌های مناسب و همچنین روابط بین آنها از درون متن اطلاق می‌شود. به عبارت دیگر استخراج اطلاعات فرآیندی برای پردازش داده‌های غیرساخت یافته (مانند متن، تصویر، صوت) یا نیمه‌ساخت یافته (مانند صفحات وب) و تبدیل آنها به مجموعه داده ساخت یافته است.
- پردازش زبان طبیعی^۶: پردازش سطح پایین و درک زبان و متن توسط کامپیوتر است.
- استخراج مفاهیم^۷: فرایند گروه‌بندی کلمات و عبارات‌های متن درون گروه‌های مشابه معنایی است.

شکل ۲-۲ نمایی دیگر از ارتباط این موضوعات و متن کاوی را نمایش داده است.

^۱ Information retrieval

^۲ Indexing

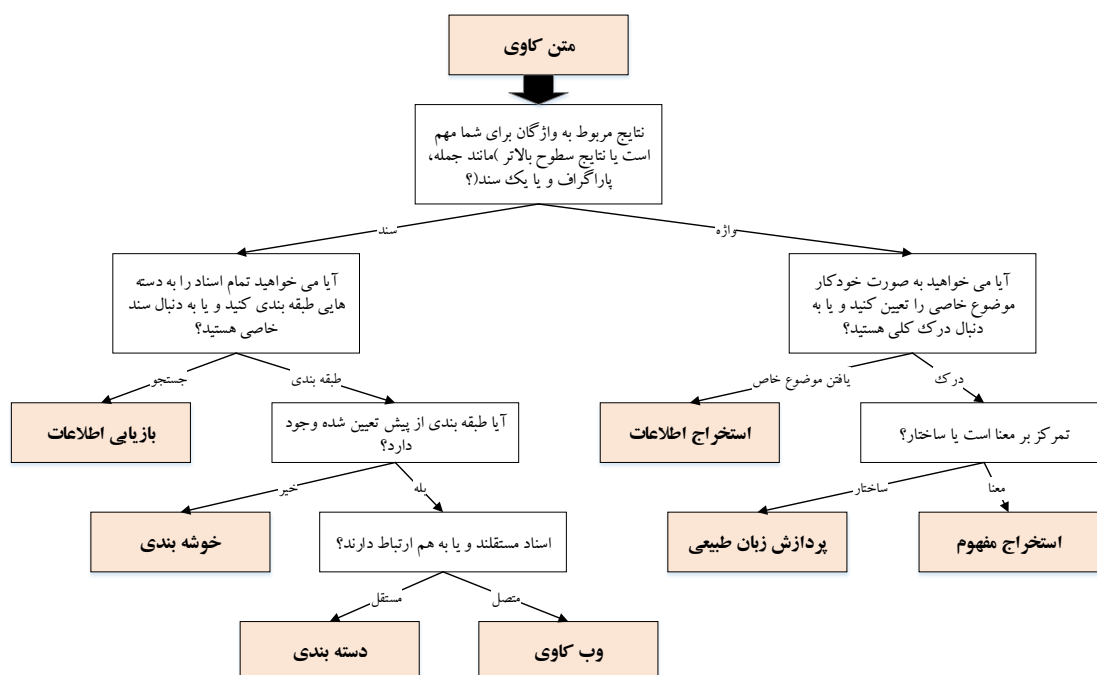
^۳ Query

^۴ Web mining

^۵ Information extraction

^۶ Natural Language Processing-NLP

^۷ Concept extraction



شکل ۲-۲- ارتباط زمینه‌های متن کاوی (FarsiYar 2020)

همچنین متن کاوی در موارد بسیار زیاد بکار می‌رود که از آن جمله می‌توان به خلاصه‌سازی اسناد متنی، خوشه‌بندی اسناد، مدل‌سازی موضوعی^۱، پردازش بین زبانی اسناد، پردازش داده‌های چندرسانه‌ای و پردازش احساسات و نظرات مشتریان اشاره نمود (Aggarwal and Zhai 2013).

درک داده‌ها و شناخت خصوصیات آنها از اصلی‌ترین نیازها در هر پردازشی است و از آنجا که داده‌ها در روش‌های متن کاوی به صورت متنی وجود دارند ضروری است تا خصوصیات آنها بررسی شوند.

۲-۳-۱ داده‌های متنی

یکی از مهمترین خصوصیات داده‌های متنی تنک^۲ و دارای ابعاد بالا بودن^۳ این داده‌هاست و تحقیقات بسیاری برای نمایش و مرتب‌سازی این داده‌ها با توجه به اهداف پردازشی صورت گرفته است (Aggarwal and Zhai 2013). از منظری دیگر، علی‌رغم اینکه درک متن و استخراج مفاهیم و معنا

^۱ Topic Modeling

^۲ Sparse

^۳ High dimensional

برای انسان‌ها کاری آسان قلمداد می‌شود اما توسعه روش‌هایی برای انتقال مفاهیم به رایانه‌ها اغلب کاری دشوار بوده و نیازمند فعالیت‌های زیر است (Zhai and Massung 2016):

- تحلیل لغوی^۱: با هدف کشف موجودیت‌های بامعنا در زبان و تعیین معنای هر موجودیت.
- تحلیل نحوی^۲: با هدف تعیین ارتباط میان واژه‌ها در یک جمله و تبیین ساختار نحوی جمله.
- تحلیل معنایی^۳: با هدف تعیین معنای یک جمله که شامل درک معنای واژه‌ها و ارتباطات نحوی آنها در یک جمله است.
- تحلیل عملی^۴: با هدف تعیین معنا در بافت. به بیانی دیگر جملات در بافت^۵ معنا می‌توانند معنایی متفاوت داشته باشند و هدف از تحلیل عملی یافتن معنای درست و واقعی جملات در بافت مربوطه است.
- تحلیل گفتمانی^۶: این تحلیل زمانی نیاز است که بخشی از یک متن (متشکل از چندین جمله) باید تحلیل شود. در این حالت، ارتباط بین جملات باید در نظر گرفته شده و با توجه به بافت تحلیل انجام شود.

برای درک بهتر موارد فوق را شکل ۲-۳ در نظر بگیرید. تحلیل لغوی در این شکل شامل تعیین نقش لغوی تمامی واژگان جمله است (تعیین اجزای گفتار^۷). تحلیل نحوی شامل تعیین نقش‌های نحوی عبارت‌ها و یا واژگان در جمله است؛ به عنوان مثال “a boy” و “the playground” عبارت اسمی محسوب می‌شوند. تحلیل معنایی شامل تعیین ارتباط عبارت‌های اسمی و فعلی است تا بتوان معنای جمله را استخراج کرد؛ اگر عبارت اسمی “a boy” به b1، عبارت اسمی “a dog” به d1 و عبارت اسمی “the playground” به p1 نگاشت شوند، ارتباط بین این عبارت‌ها که نمایانگر فعالیت chasing باشد را می‌توان به صورت chasing(d1,b1,p1) نمایش داد. تحلیل عملی این جمله می‌توان شامل این باشد که: گوینده این جمله ممکن است درخواستی داشته باشد به عنوان مثال، با صاحب سگ یادآوری کند که به دنبال سگ خود برود.

^۱ Lexical

^۲ Syntactic

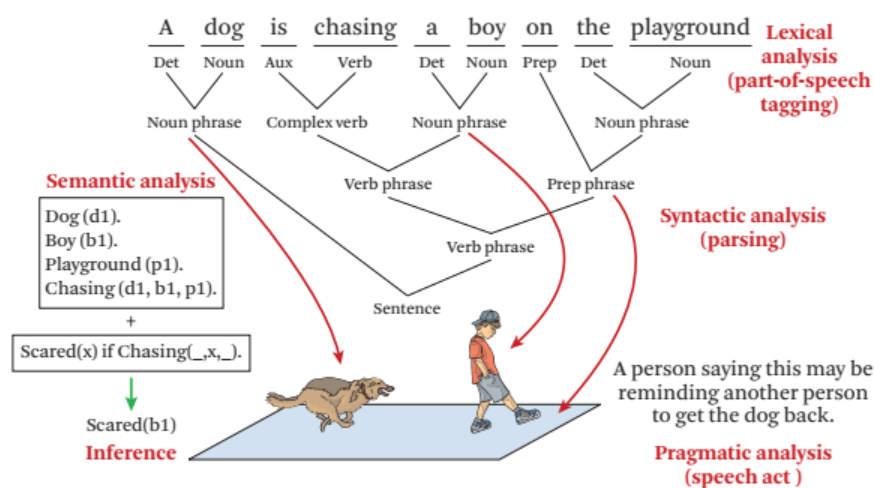
^۳ Semantic

^۴ Pragmatic

^۵ Context

^۶ Discourse

^۷ Parts of Speech- POS



شکل ۲-۳. نمونه‌ای از پردازش‌ها در درک زبان طبیعی (Zhai and Massung 2016)

۲-۳-۲ نمایش داده‌های متنی

یکی از راه‌ها برای پردازش داده‌های متنی، تبدیل آنها به بردارهایی عددی است تا از این طریق بتوان محاسبات مورد نیاز را انجام داد. روش‌های متعدد و کاربردی برای این تبدیل در ادبیات وجود دارد که از جمله آنها می‌توان به موارد زیر اشاره کرد:

□ مدل فضای برداری^۱: مدل فضای برداری یکی از اولین روش‌ها برای تبدیل واژگان یک متن به بردارهای عددی است که در آن ساختار نحوی و معنایی واژگان در نظر گرفته نمی‌شود. در برداری عددی در این روش براساس اهمیت واژه تشکیل شده و یکی از پرکاربردترین رویه‌ها برای استخراج این بردار استفاده از روش فراوانی واژه-فراوانی معکوس سند (TF-IDF)^۲ (رابطه (۱-۲)) است (Aggarwal and Zhai 2013).

$$tf-idf(w) = tf \times \log \frac{N}{df(w)} \quad (1-2)$$

در این رابطه $tf(w)$ نشانگر فراوانی واژه (تعداد تکرار واژه در یک سند)، $df(w)$ نشانگر فراوانی سند (تعداد اسنادی که حاوی واژه مورد نظر هستند) و N نشانگر تعداد اسناد مورد بررسی در پیکره^۳ و یا پایگاه داده مورد بررسی است (Aggarwal and Zhai 2013). رابطه (۱-۲) در ادبیات به صورت‌های دیگری نیز تعریف شده است. مشکل اصلی این روش وجود محاسبه بردارهایی تنک با ابعاد بسیار زیاد است.

□ مدل بازنمایی کلمات^۴: هدف اصلی در این مدل نمایش واژگان با بردارهایی با ابعاد کم و با توجه به بافت است. به بیانی دیگر، در این مدل ارتباط معنایی بین واژگان حفظ شده و به دلیل ابعاد کمتر نسبت به روش TF-IDF پیچیدگی محاسباتی عموماً کمتر است (Naili, Chaibi, and Ben Ghezala 2017). همچنین در این مدل، بردار واژگان با معنای نزدیک به هم، به یکدیگر نزدیک‌تر هستند (فاصله کمتری نسبت به سایر واژه‌ها دارند) (Toshevska, Stojanovska, and Kalajdjieski 2020). روش‌های متعددی در ذیل این مدل طبقه‌بندی می‌شوند که از آن میان می‌توان به مدل‌های تحلیل معنای نهفته^۵ (Mikolov Word2Vec)

^۱ Vector space model

^۲ Term Frequency-Inverse Document Frequency- TF-IDF

^۳ Corpus

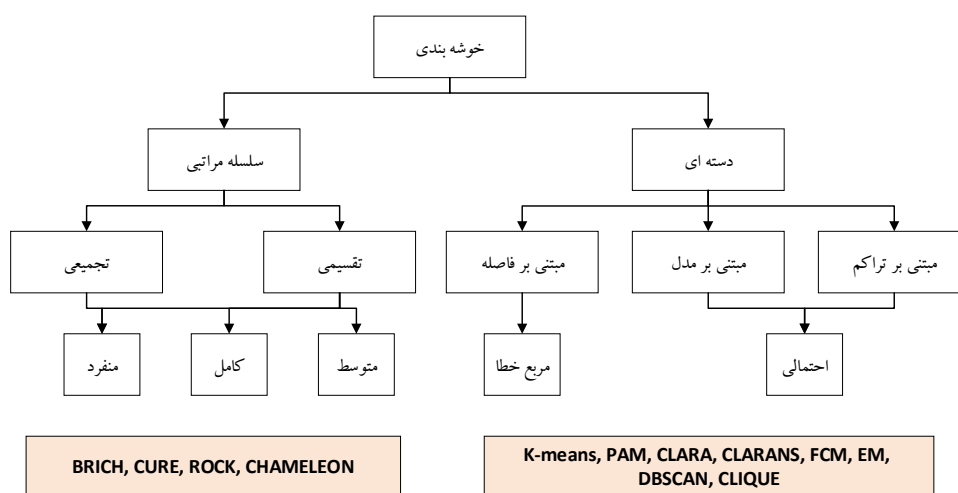
^۴ Word embedding

^۵ Latent Semantic Analysis- LSA

(Peters ELMO, (Pennington, Socher, and Manning 2014) GloVe, et al. 2013)
(Devlin et al. 2019) BERT, et al. 2017) اشاره نمود.

۴-۲ خوشه‌بندی

خوشه‌بندی داده‌ها به معنای یافتن گروه‌هایی از موارد مشابه است به صورتی که اعضای یک خوشه تا حد ممکن یکدیگر شبیه باشند و اعضای خوشه‌های مختلف تا حد ممکن متفاوت باشند. برای انجام این امر الگوریتم‌های متعدد خوشه‌بندی توسعه یافته‌اند که می‌توان آنها را در دو دسته کلی طبقه‌بندی کرد: خوشه‌بندی سلسله‌مراتبی^۱ و خوشه‌بندی دسته‌ای^۲ (شکل ۴-۲) (Saxena et al. 2017).



شکل ۴-۲- طبقه‌بندی الگوریتم‌های خوشه‌بندی (Saxena et al. 2017)

در خوشه‌بندی سلسله‌مراتبی، خوشه‌ها به وسیله تقسیم کردن الگوها به زیرشاخه‌هایی مشابه و یا تجمیع زیرشاخه‌ها به سرشاخه‌هایی مرتبط بدست می‌آیند. خوشه‌بندی تجمیعی^۳ و خوشه‌بندی تقسیمی^۴ دو زیرمجموعه کلان این نوع خوشه‌بندی هستند. خوشه‌بندی تجمیعی رویکردی پایین به بالا دارند و از یک عضو در هر خوشه شروع می‌کنند. به بیانی دیگر، ابتدا تعداد خوشه‌ها برابر با تعداد داده‌ها است و در طی مراحل، این تعداد به یک خوشه با دربرگیری تمام داده‌ها تبدیل می‌شود مگر آنکه شرایط ختم الگوریتم محقق شود. این روش برای خوشه‌بندی تقسیمی معکوس است. یعنی ابتدا همه داده‌ها در یک خوشه فرض شده و آنقدر مراحل تکرار می‌شوند تا در انتها هر داده در یک خوشه قرار گیرد

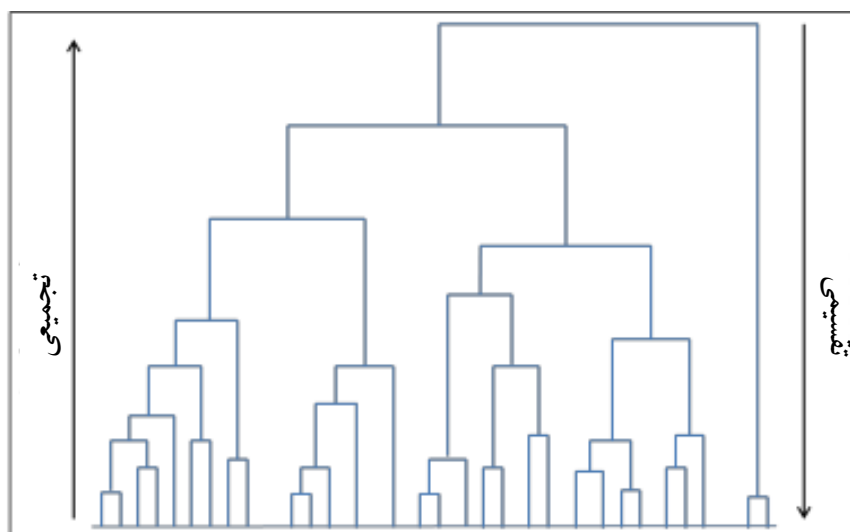
^۱ Hierarchical

^۲ Partitional

^۳ Agglomerative

^۴ Divisive

مگر آنکه شرایط ختم الگوریتم محقق شود (شکل ۲-۵). الگوریتم‌هایی چون ^۱BIRCH، ^۲CURE، ROCK و CHAMELEON از نمونه‌های خوشه‌بندی سلسله‌مراتبی هستند (Saxena et al. 2017).



شکل ۲-۵- خوشه‌بندی تجمیعی و خوشه‌بندی تقسیمی (Saxena et al. 2017)

مقابل خوشه‌بندی سلسله‌مراتبی، خوشه‌بندی دسته‌ای قرار دارد. در این خوشه‌بندی داده‌ها به k خوشه بدون توجه به ساختار سلسله‌مراتبی و با هدف بهینه ساختن تابع هدفی مشخص خوشه‌بندی می‌شوند. تابع مشابهت در این روش می‌تواند براساس فاصله و یا براساس مدل‌های احتمالی و یا تکاملی تعریف شوند. الگوریتم‌هایی چون k -means، Fuzzy C-means و DBSCAN از جمله الگوریتم‌های پرکاربرد در این دسته هستند (Saxena et al. 2017).

انتخاب و بکارگیری هریک از این روش‌ها با توجه به معیارهای متعددی صورت می‌گیرد که از آن جمله می‌توان به میزان پیچیدگی محاسباتی^۳ هر روش، امکان پردازش نوع داده‌های دردست بررسی (ساخت یافته، غیرساخت یافته و یا نیمه‌ساخت یافته، دودویی، احتمالی، عدد صحیح و ...)، امکان پردازش تعداد داده‌های دردست بررسی (حجم و یا کلان)، امکان پردازش خصوصیات داده‌های دردست بررسی (با ابعاد کم و یا زیاد)، و یا میزان حساسیت روش به اغتشاشات داده‌ها اشاره نمود (Saxena et al. 2017).

^۱ Balanced Iterative Reducing and Clustering Using Hierarchies-BIRCH

^۲ Clustering Using Representatives- CURE

^۳ Complexity

همچنین در هر روشی که انتخاب می‌شود گام‌های زیر طی خواهد شد (طی همه گام‌ها ضروری نیست) (Feldman and Sanger 2007):

- گام ۱: بازنمایی مسئله شامل استخراج و یا انتخاب ویژگی‌های داده‌ها
- گام ۲: انتخاب و تعریف معیار شباهت و نزدیکی در خوشه‌بندی
- گام ۳: خوشه‌بندی داده‌ها
- گام ۴: خلاصه‌سازی نتایج و داده‌ها
- گام ۵: ارزیابی

۵-۲ خوشه‌بندی اسناد

با این تعریف می‌توان گفت خوشه‌بندی در متن کاوی کاربردهای مهمی به ویژه در مرتب‌سازی اسناد با هدف بهبود بازیابی و ارتقاء نتایج جستجو خواهد داشت (Aggarwal and Zhai 2013) چرا که با کمک آن می‌توان نتایج جستجو را به صورت دیداری به کاربر نمایش داده و یا اسناد مشابه را در خوشه‌های مشابه به کاربر ارائه داد. همچنین، از خوشه‌بندی خود کار اسناد می‌توان به عنوان ابزاری برای یافتن موضوعات کلان اسناد پایگاه داده استفاده نمود که با کمک آن کاربران قادرند اسناد مورد نیاز خود را بدون جستجو و صرفاً از طریق مرور موضوعات کلان بیابند (این امر زمانی مهم است که کاربر آشنایی چندانی با موضوع مورد جستجوی خود ندارد و نیازمند یافتن سرخ‌هایی برای جستجوی عمیق‌تر است). از منظر بازیابی اطلاعات نیز، از خوشه‌بندی می‌توان به عنوان ابزاری برای بهبود کارایی سیستم‌های جستجو استفاده کرد. به بیانی دیگر، می‌توان اسناد پایگاه داده را از قبل خوشه‌بندی کرد و در هنگام جستجوی کاربر به جای نمایش منفرد نتایج، خوشه‌های بازیابی شده با بیشترین شباهت به عبارت پرس‌وجو را نمایش داد (Aljaber et al. 2010). متن کاوی دارای ابزارهای تجاری متعددی است که در بسیاری از آنها امکان خوشه‌بندی اسناد وجود دارد. ^۱Lemur و ^۲BOW دو نمونه از این ابزارها هستند (Aggarwal and Zhai 2013).

گام‌های خوشه‌بندی متن با توجه به آنچه در بخش پیش مطرح شد به شرح زیر است:

۵-۲-۱ گام ۱: بازنمایی مسئله

^۱ <http://www.lemurproject.org/>

^۲ <http://www.cs.cmu.edu/~mccallum/bow/>

اسناد متنی را می‌توان یا به صورت داده‌های دودویی (وجود و یا عدم وجود واژه‌ای در متن) و یا به صورت متن (با سطح ریزدانگی^۱ حرف، واژه، جمله، پاراگراف، تک صفحه، بخش و یا کل سند) بازنمایی کرد (Aggarwal and Zhai 2013).

در حالت بازنمایی دودویی، وجود و یا عدم وجود واژه‌ای در متن بررسی می‌شود و تکرار آن واژه در یک سند شمرده خواهد شد یا حتی ممکن است ضریب وزنی برای واژگان متن، با توجه به اهمیت آن واژه، در نظر گرفته شود (مانند بازنمایی TF-IDF) (Aggarwal and Zhai 2013). الگوریتم‌های متعددی در ادبیات از این بازنمایی استفاده می‌کنند که از آن میان می‌توان به (Andritsos et al. 2004)، (Gibson, Kleinberg, and Raghavan 2000) و (Guha, Rastogi, and Shim 2001) اشاره نمود. با این حال از این بازنمایی نمی‌توان به صورت کارا برای خوشه‌بندی اسناد استفاده کرد چرا که یک سند متنی خصوصیات ویژه‌ای دارد (Aggarwal and Zhai 2013):

- ابعاد داده‌های بدست آمده از این روش بازنمایی بسیار زیاد بوده و تنگ است. این امر در مورد اسناد متنی کوتاه بسیار شدیدتر است.
- با اینکه تعداد واژگان مجموعه داده زیاد است اما واژه‌ها اغلب با یکدیگر مرتبط هستند و این بدین معنی است که تعداد مفاهیم و یا مولفه‌های اساسی در داده‌ها بسیار کمتر از تعداد ویژگی‌هاست.
- تعداد واژگان در اسناد مختلف متفاوت است و لذا ضروری است تا اسناد نرمالسازی شوند.

بر این مبنا روش‌هایی چون استفاده از فراوانی معکوس سند، قدرت یک واژه (محاسبه میزان اطلاع دهندگی یک واژه در افتراق اسناد) و رتبه‌دهی بر مبنای آن‌تروپی توسعه داده شده‌اند (Aggarwal and Zhai 2013).

در روش رتبه‌دهی بر مبنای آن‌تروپی، کیفیت یک واژه براساس کاهش آن‌تروپی زمانی که آن واژه از پیکره حذف می‌شود محاسبه خواهد شد. اگر $E(t)$ نشان‌دهنده آن‌تروپی واژه t در مجموعه‌ای متشکل از N سند باشد و $S_{ij} \in (0,1)$ شباهت i th و j th سند را بعد از حذف واژه t نشان دهد، آنگاه (Aggarwal and Zhai 2013):

$$E(t) = -\sum_{i=1}^n \sum_{j=1}^n (S_{ij} \log(S_{ij}) + (1-S_{ij}) \log(1-S_{ij})) \quad (2-2)$$

^۱ Granularity

$$S_{ij} = 2 \frac{\text{dist}(i,j)}{\text{dist}} \quad (3-2)$$

که $\text{dist}(i,j)$ فاصله i th و j th سند بعد از حذف واژه t و $\overline{\text{dist}}$ میانگین فاصله بین تمامی اسناد بعد از حذف واژه t است. پیچیدگی محاسباتی این روش $O(n^2)$ بوده و بنابراین در پیکره‌های بزرگ کارایی زیادی نخواهد داشت (Aggarwal and Zhai 2013).

دسته دیگری از روش‌ها بر کاهش ابعاد داده‌ها تمرکز دارند که از آن میان می‌توان به روش تحلیل مولفه‌های اساسی^۱، تجزیه مقادیر منفرد^۲ و یا تجزیه نامنفی ماتریسی^۳ اشاره کرد (Aggarwal and Zhai 2013).

یکی از کاستی‌های مشترک در بین روش‌های اشاره شده آن است در این روش‌ها ارتباط بین واژگان در نظر گرفته نشده و برای پردازش داده‌های حجیم و یا کلان بهینه نیستند. در سال‌های اخیر روش‌های جدیدتری براساس یادگیری ماشین و نیز شبکه‌های عمیق توسعه یافته و کارایی خود را در پردازش کردن داده‌های حجیم متنی اثبات کرده‌اند. به برخی از این روش‌ها در بخش متن کاوی (0:نمایش داده‌های متنی) به آن اشاره شد. به جهت اهمیت در این بخش به دو روش مهم در این حوزه، word2vec و BERT می‌پردازیم.

Word2vec

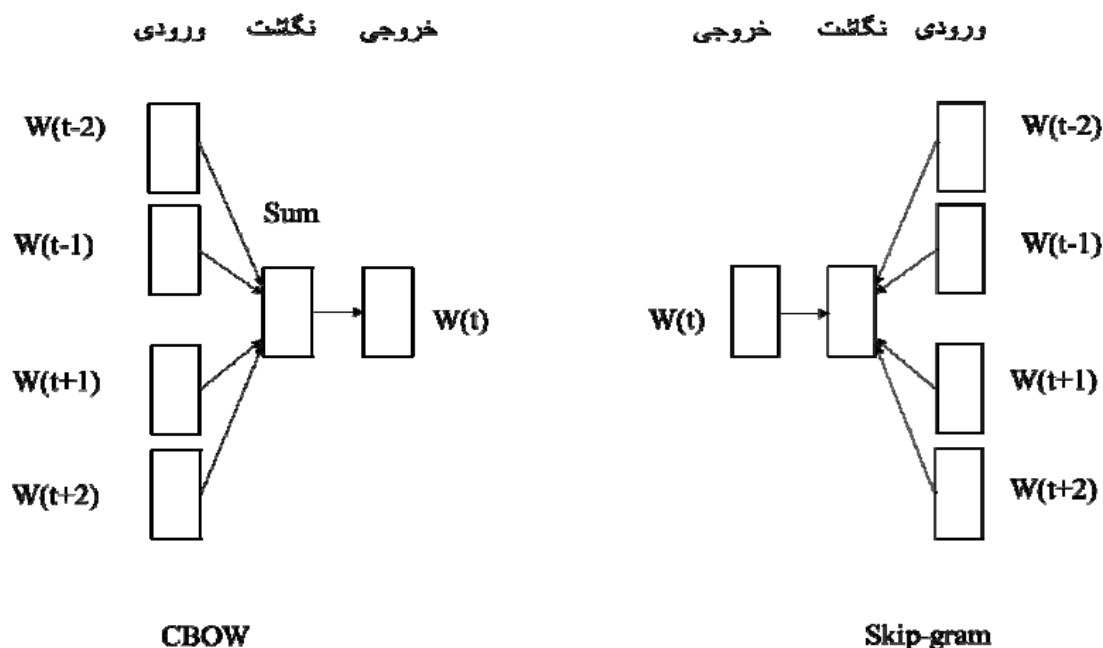
یکی از مهمترین الگوریتم‌های بازنمایی کلمه، روش word2vec است که در سال ۲۰۱۳ توسط میکولوف و همکاران (Mikolov et al. 2013) برای نمایش برداری کلمات در فضای ابعادی کم پیشنهاد شد. این روش براساس یک شبکه عصبی با یک لایه پنهان طراحی شده که کلمات را به عنوان بردارهایی در یک فضای برداری چند بعدی نمایش می‌دهد. درواقع این روش، یک مدل پیش‌بینی را با آموزش شبکه عصبی روی مجموعه‌ای عظیمی از اسناد متنی می‌سازد که براساس آن می‌تواند پیش‌بینی کند، هر کلمه چه میزان به کلمه دیگری ارتباط دارد. مدل آموزش داده شده، روابط معنایی بین کلمات را با توجه به نحوه قرارگیری آنها در پیکره متنی استخراج می‌کند و دقت مدل به تعداد دفعات هم‌رخدادی کلمات درون یک متن در پیکره متنی وابسته است. بنابراین هر چه حجم داده‌های پیکره بیشتر باشد، دقت این روش افزایش خواهد یافت (محبی، جلالی منش، ۱۳۹۷).

^۱ Principal Component Analysis- PCA

^۲ Singular Value Decomposition- SVD

^۳ non-negative matrix factorization- NMF

برای آموزش شبکه عصبی از دو مدل استفاده می‌شود: مدل کیسه لغات پیوسته^۱ (CBOW) و مدل Skip-gram. در شکل ۶-۲ ساختار دو مدل CBOW و Skip-gram نمایش داده شده است و $w(t)$ کلمه هدف^۲، N اندازه پنجره و کلمات همسایگی از $w(t - N)$ تا $w(t + N)$ هستند (محبی، جلالی‌منش، ۱۳۹۷).



شکل ۶-۲- ساختار دو مدل CBOW و Skip-gram (محبی، جلالی‌منش، ۱۳۹۷)

در مدل CBOW شبکه عصبی آموزش داده می‌شود که با یک اندازه پنجره مشخص و با استفاده از واژگان همسایه یا بافت^۳، کلمه هدف را پیش‌بینی می‌کند. بنابراین در این مدل برای آموزش شبکه، همسایگی یک کلمه هدف به عنوان ورودی به شبکه عصبی داده شده و وزن‌ها براساس پیکره متنی و یا پایگاه داده، تعیین می‌شوند. در مدل Skip-gram شبکه عصبی به گونه‌ای آموزش داده می‌شود تا با یک اندازه پنجره مشخص و با استفاده از یک کلمه هدف، کلمه‌های همسایه را پیش‌بینی کند (محبی، جلالی‌منش، ۱۳۹۷).

در CBOW قبل از پیش‌بینی کلمه هدف، از بردارهای کلمات بافت میانگین گرفته می‌شود. اما در Skip-gram عمل میانگین‌گیری کلمات انجام نمی‌شود. لذا Skip-gram در مواقعی که کلمات خاص

^۱ Continuous Bag of Words- CBOW

^۲ Target Word

^۳ Context

و کم تکرار در پیکره زیاد است، بهتر عمل می‌کند. در هر دو مدل، آموزش با توجه به یک پنجره از کلمات همسایگی انجام می‌شود و همچنین هنگامی که آموزش همگرا شد، بردار خروجی نادیده گرفته شده و تنها وزن موجود بین لایه پنهان و لایه خروجی به عنوان نمایشگر برداری کلمات استفاده می‌شود (محبی، جلالی منش، ۱۳۹۷).

در CBOW برای پیش بینی کلمه هدف، بردارهای کلمات موجود در پنجره همسایگی تجمیع و مورد استفاده قرار می‌گیرد. بنابراین، اگر پیکره متنی شامل کلمات $w_1, w_2, \dots, w_T$ و اندازه پنجره برابر C باشد، تابع هدف برابر است با (در این تابع احتمال رخداد کلمه w_t با توجه به کلمات همسایگی آن با پنجره به طول C محاسبه می‌شود) (محبی، جلالی منش، ۱۳۹۷):

$$\frac{1}{T} \sum_{t=1}^T \log p \left(w_t \mid \sum_{-C \leq j \leq C, j \neq 0} w_{t+j} \right) \quad (۴-۲)$$

در مدل Skip-gram کلمات پنجره به طور مستقل برای کلمه هدف، پیش‌بینی می‌شوند. بنابراین تابع هدف برابر است با (محبی، جلالی منش، ۱۳۹۷):

$$\frac{1}{T} \sum_{t=1}^T \log p \left(w_{t+1} \mid \sum_{-C \leq j \leq C, j \neq 0} w_t \right) \quad (۵-۲)$$

برای محاسبه احتمال $p(w_{t+j} | w_t)$ می‌توان از تابع Softmax استفاده کرد (محبی، جلالی منش، ۱۳۹۷):

$$p(w_o | w_i) = \frac{\exp(v_{w_o}'^T v_{w_i})}{\sum_{w=1}^W \exp(v_w'^T v_{w_i})} \quad (۶-۲)$$

که v_w, v_w' بازنمایی برداری کلمه w هستند و W تعداد کل کلمات است. این رابطه پیچیدگی محاسباتی بالایی دارد چرا که تعداد کلمات یک پیکره بزرگ که معمولاً برای آموزش می‌شود، بسیار زیاد است و محاسبات آن زمانبر خواهد بود. به همین منظور دو استراتژی سافت مکس سلسله‌مراتبی^۱ و نمونه‌برداری منفی^۲ پیشنهاد شد (محبی، جلالی منش، ۱۳۹۷).

پارامترهای آموزش این مدل شامل موارد زیر است (محبی، جلالی منش، ۱۳۹۷):

^۱ Hierarchical softmax

^۲ Negative sampling

- اندازه پنجره: افزایش اندازه پنجره ممکن است شامل کلماتی شود که با کلمه هدف مرتبط نیستند و در نتیجه زمان آموزش مدل را طولانی‌تر و کاهش اندازه پنجره منجر به حضور ایست‌واژه می‌شود. بنابراین انتخاب صحیح اندازه پنجره تاثیر بسزایی در عملکرد مدل دارد.
- ابعاد بردار: تعداد ابعاد بردار، اندازه لایه نگاشت (میانی) را نشان می‌دهد.
- نمونه‌برداری: این پارامتر برای حذف واژگانی که فراوانی بیشتری از حد آستانه مشخص شده دارند، تعیین می‌شود. به عبارت دیگر، واژگان پرتکرار، همانند ایست‌واژه^۱، بار معنایی برای استخراج روابط معنایی بین واژگان ندارند. احتمال این واژگان به صورت زیر محاسبه می‌شود t حد آستانه و $f(w_t)$ فراوانی واژه در کل پیکره یا پایگاه داده است):

$$p(w_t) = 1 - \sqrt{\frac{t}{f(w_t)}} \quad (7-2)$$

- کمترین فراوانی^۲: این پارامتر به حذف واژگانی می‌پردازد که کمتر از تعداد خاصی در کل پیکره متنی ظاهر شده‌اند.
- نمونه‌برداری منفی: هدف از نمونه‌برداری منفی، افزایش سرعت زمان آموزش است. به عبارت دیگر، در هرتکرار و بروزرسانی وزن‌های شبکه، به جای به‌روزرسانی بردارهای تمام کلمات، تنها بخشی از کلمات به‌روزرسانی می‌شود.

BERT

روش BERT^۳ در سال ۲۰۱۹ توسط دولین و همکاران (Devlin et al. 2019) معرفی شد و برخلاف مدل‌های پیشین برای پیش آموزش نمایش دوسویه شبکه عمیق با کمک داده‌های بدون برچسب طراحی شده است. این مدل از دو گام اصلی تشکیل شده است: پیش آموزش^۴ و بهینه‌سازی^۵. در طی پیش آموزش، مدل برای داده‌های بدون برچسب آموزش داده می‌شود و برای بهینه‌سازی مدل آموزش داده شده، از داده‌های با برچسب استفاده می‌شود. مدل مورد استفاده، مدل چندلایه دوسویه ترنسفورمر بر اساس پژوهش (Vaswani et al. 2017) توسعه یافت است. ایده اصلی تقسیم‌بندی فرآیند به دو مرحله به این دلیل است که به طور معمول، پایگاه داده مسئله مورد نظر به اندازه کافی بزرگ نیست

^۱ Stop words

^۲ Min-count

^۳ Bidirectional Encoder Representations from Transformers- BERT

^۴ Pre-training

^۵ Fine-tuning

که بتواند پشتیبانی مناسبی از فرآیند آموزش شبکه عمیق داشته باشد. به همین دلیل در طی این مدل و برای آموزش زبانی ابتدا از پیکره‌های متنی بزرگ استفاده شده و سپس مدل ساخته شده برای مسئله مورد نظر تطبیق می‌یابد.

یکی از اجزای اصلی معماری ترنسفورمر مکانیزم خودتوجه^۱ است که در آن اجزای توالی ورودی و خروجی ارزیابی و وزن‌دهی می‌شوند و بر همین اساس می‌تواند روابط بین اجزا را حتی در جملات با وابستگی‌های طولانی، مدل‌سازی کند (هوشینو، ۱۴۰۰).

در یک مثال فرض کنید می‌خواهیم یک بازنمایی از جمله «تام یک گربه سیاه است» ایجاد کنیم. ممکن است BERT توجه خود را به واژه «تام» معطوف کند و به صورت همزمان کلمه «گربه» را رمزگذاری کند و توجه کمتری به لغات «است»، «یک» و «سیاه» نشان دهد. این روند را می‌توان به صورت برداری از وزن‌ها نشان داد (برای هر یک از واژگان تشکیل‌دهنده جمله). پس از اینکه مدل، هر یک از کلمات تشکیل‌دهنده توالی را کدگذاری کرد، بردارها محاسبه می‌شوند و در نتیجه یک ماتریس مربعی به دست می‌آید که نگاشت خودتوجه نامیده می‌شود. بدیهی است که ارتباط میان «تام» و «گربه» همیشه بهترین رابطه نیست. مدل برای پاسخ دادن به سؤالات مرتبط با رنگ گربه بهتر است به جای «تام» بر روی کلمه «سیاه» تمرکز کند. خوشبختانه در چنین مواردی مدل مجبور به انتخاب نیست. برتری BERT (و دیگر ترنسفورمرها) تا حد زیادی ناشی از این واقعیت است که لایه‌های این مدل از شاخه‌های^۲ متعددی تشکیل شده و تمامی این شاخه‌ها شیوه ساخت نگاشت‌های خود توجه مستقل را یاد می‌گیرند. به لحاظ نظری، این امر مدل را قادر می‌سازد تا «در موقعیت‌های گوناگون به اطلاعات به دست آمده از زیرفضای بازنمایی‌های مختلف توجه کند». به زبان ساده، در این حالت مدل برای حل مسئله مورد بحث، از میان بازنمایی مختلف حق انتخاب دارد (هوشینو، ۱۴۰۰).

۲-۵-۲ گام ۲: معیار شباهت

معیار شباهت معیاری برای سنجش میزان شباهت اجزای درون یک خوشه بوده و نقش مهمی در کارایی خوشه‌بندی دارد. این معیار بایستی بر اساس ویژگی‌های داده‌ها و خصوصیات روش خوشه‌بندی انتخاب شود. معیار فاصله یکی از معیارهای اصلی برای محاسبه شباهت بین دو داده است و انواع مختلفی از معیار فاصله در ادبیات توسعه داده شده است. چهار ویژگی ضروری زیر فصل

^۱ Self-attention

^۲ Head

مشترک معیار فاصله بین دو داده x و y ($D(x, y)$) است (Zarinbal, Fazel Zarandi, and Turksen 2014):

۱. $D(x, y) \geq 0$: فاصله بین دو داده x و y در همه شرایط نامنفی است.
۲. $x = y \leftrightarrow D(x, y) = 0$: اگر دو داده یکی باشند فاصله بین آنها صفر است و برعکس.
۳. $D(x, y) = D(y, x)$: فاصله متقارن است.
۴. $D(x, z) \leq D(x, y) + D(y, z)$: فاصله خاصیت تحدب دارد.

در ادامه به برخی از مهم‌ترین و پرکاربردترین معیارهای فاصله اشاره می‌شود.

□ فاصله مینکویسکی^۱: روش محاسبه این فاصله در رابطه (۲-۱) تعریف شده است.

$$D_p(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_k (x_{ik} - x_{jk})^p \right)^{1/p} \quad (۲-۸)$$

در این رابطه اگر $p = 2$ باشد فاصله اقلیدسی بدست می‌آید.

□ فاصله کسینوسی^۲: این فاصله، زاویه بین دو بردار را محاسبه کرده و به علت بهره‌گیری از فضای برداری در متن کاوی، در این حوزه بسیار بکار رفته است (رابطه (۲-۹) (Aggarwal and Zhai 2013):

$$D_{\cosine}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|} = \frac{\sum_i x_i y_i}{\sqrt{\sum_i (x_i)^2} \sqrt{\sum_i (y_i)^2}} \quad (۲-۹)$$

□ فاصله ژاکارد^۳: این فاصله برای محاسبه فاصله دو مجموعه از هم محاسبه می‌شود. به بیانی دیگر، در این معیار وجود و یا عدم وجود عناصر در مجموعه‌ها به عنوان فاصله تعریف شده است (رابطه (۲-۱۰) (Zhai and Massung 2016):

$$D_{Jaccard}(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} \quad (۲-۱۰)$$

^۱ Minkowski

^۲ Cosine

^۳ Jaccard

□ فاصله آنتروپی نسبی^۱ یا واگرایی^۲ Kullback-Leibler: این معیار فاصله بیشتر در تئوری اطلاعات استفاده شده است و با کمک آن شباهت دو داده یا دو سند بر اساس توزیع احتمالی آن‌ها و یا میزان اطلاعات تبادل شده بین آن دو محاسبه می‌شود. دو توزیع P و Q را در نظر بگیرید؛ فاصله KL این دو توزیع از رابطه (۱۱-۲) بدست می‌آید (Zarinbal, Fazel (2014): Zarandi, and Turksen

$$D(P \parallel Q) = \sum_x p(x) \ln \frac{p(x)}{q(x)} \quad (11-2)$$

همچنین از آنجا که این فاصله مقارن نیست، در محاسبات عموماً از میانگین آن استفاده می‌شود (رابطه (۱۲-۲)) (Huang 2008).

$$D_{AvgKL}(P \parallel Q) = \pi_1 D_{kl}(P \parallel M) + \pi_2 D_{kl}(Q \parallel M) \quad (12-2)$$

$$\text{که در آن } M = \pi_1 P + \pi_2 Q, \pi_1 = \frac{P}{P+Q}, \pi_2 = \frac{Q}{P+Q}$$

۲-۵-۳ گام ۳: خوشه‌بندی داده‌ها

در بخش قبل به روش‌های خوشه‌بندی اشاره شد. در حوزه خوشه‌بندی متن به صورت معمول از روش‌های زیر استفاده می‌شود:

- Single-link agglomerative clustering
- Complete-link agglomerative clustering
- Expectation-maximization
- K-means

همچنین نحوه بررسی اسناد در خوشه‌بندی در این بخش به رویکردهای خوشه‌بندی در رابطه با داده‌های متنی پرداخته می‌شود.

خوشه‌بندی براساس واژگان و عبارت‌ها

در این رویکرد، ابتدا واژگان خوشه‌بندی شده و از مهم‌ترین خوشه‌ها برای یافتن اسناد مرتبط استفاده می‌شود. اگر پیکره حاوی n و d سند باشد، ماتریس واژه-سند، ماتریسی با $n \times d$ بعد است که در آن $(i, j)^{th}$ ورودی تکرار i^{th} واژه در j^{th} اسناد است.

^۱ Relative entropy

^۲ Divergence

$$\text{term - document matrix} = \begin{bmatrix} X_{11} & \dots & X_{1d} \\ X_{21} & \dots & X_{2d} \\ \vdots & \ddots & \vdots \\ X_{n1} & \dots & X_{nd} \end{bmatrix}$$

اگر خوشه‌بندی بر مبنای سطرهای این ماتریس انجام شود، خوشه‌بندی واژگان انجام شده است و اگر خوشه‌بندی بر مبنای ستون‌های این ماتریس انجام شود، خوشه‌بندی اسناد انجام شده است. این دو رویکرد بسیار نزدیک به هم بوده و می‌توان از هر دو روش همزمان استفاده کرد که در ادبیات از آن با خوشه‌بندی همزمان^۱ یاد می‌شود. خوشه‌بندی زیرمجموعه‌ای^۲ یا خوشه‌بندی نگاشت شده^۳ نیز از جمله دیگر رویکردهای خوشه‌بندی در این حوزه است. در برخی دیگر از روش‌ها به ترتیب واژگان توجه شده است. به بیانی دیگر، از عبارت‌های یک سند به عنوان داده‌های خوشه‌بندی استفاده می‌شود (Aggarwal and Zhai 2013).

خوشه‌بندی براساس موضوع

مدلسازی موضوعی^۴ یکی از رویکردهای پرکاربرد در خوشه‌بندی احتمالی اسناد است که در آن برای اسناد موجود در پیکره مدل مولد احتمالی^۵ ساخته می‌شود. در این روش فرض می‌شود که هر سند با احتمالی به یکی از k موضوع پیکره تعلق دارد و بنابراین یک سند می‌تواند به صورت همزمان به چند موضوع تعلق داشته باشد. همچنین فرض می‌شود که هر موضوع شامل برداری احتمالی است که احتمال واژگان مربوط به آن موضوع را گزارش می‌کند (Aggarwal and Zhai 2013).

خوشه‌بندی نیمه نظارتی

در برخی از کاربردها، اطلاعاتی قبلی از تعداد خوشه‌ها و یا خصوصیات خوشه‌ها وجود دارد که از آن می‌توان به عنوان برچسب برای برخی داده‌ها و یا برخی موضوعات داده‌ها استفاده کرد. چنین دانشی ممکن است در بهبود نتایج خوشه‌بندی موثر باشد (Aggarwal and Zhai 2013). به عنوان نمونه، در مدیریت محتوا، هدف دسته‌بندی خودکار اسناد متنی است که تعداد آنها اغلب بسیار زیاد است (چندین میلیون). در این حالت وجود اطلاعات قبلی از اسناد، هم‌رخدادی آنها و یا وجود اختلاف زیاد بین اسناد می‌تواند بر سرعت و کارایی خوشه‌بندی بیافزاید (Srivastava and Sahami 2009). چنین

^۱ Co-clustering

^۲ Subspace clustering

^۳ Projected clustering

^۴ Topic Modeling

^۵ Probabilistic generative model

رویکردی، خوشه‌بندی نیمه‌نظارتی نامیده شده و رویکردی بین خوشه‌بندی و دسته‌بندی^۱ است چرا که از ساختارهای ضمنی داده‌ها استفاده می‌کند اما مقید به آنها نیست (Aggarwal and Zhai 2013). روش‌های مرسوم در این خوشه‌بندی به دو دسته روش‌های مبتنی بر محدودیت^۲ و یا روش‌های مبتنی بر فاصله^۳ دسته‌بندی می‌شوند. در روش‌های مبتنی بر محدودیت، الگوریتم خوشه‌بندی در هر مرحله بر اساس برجسب‌ها و یا محدودیت‌های تعریف شده خود را اصلاح می‌کند، به عنوان نمونه اگر طبق محدودیت‌های موجود دو داده باید در یک خوشه باشند (must-link) و دو داده نباید در یک خوشه قرار بگیرند (cannot-link)، الگوریتم جریمه و پاداش متناظر برای نتایج در نظر گرفته و حرکت خود را به سمت بیشترین سود اصلاح می‌کند. الگوریتم K-means و توسعه‌های آن نظیر COP-Kmeans، Constrained Vector Quantization Error (CVQE) و linear-time CVQE (LCVQE) از عمده الگوریتم‌های مورد استفاده در این رویکرد هستند (Srivastava and Sahami 2009).

خوشه‌بندی عمیق

در صورتی که داده‌های مورد پردازش دارای ابعاد زیاد باشند، مانند متن یا تصویر، الگوریتم‌های خوشه‌بندی دقت خوبی نخواهند داشت و نمی‌توانند داده‌ها را به درستی خوشه‌بندی کنند. برای حل این مشکل دو رویکرد وجود دارد (Min et al. 2018):

□ کاهش بعد داده‌ها با کمک روش‌هایی مانند PCA^۴ و سپس بکارگیری روش‌های معمول

خوشه‌بندی مانند K-means

□ بکارگیری خوشه‌بندی عمیق^۵. این رویکرد زیرمجموعه یادگیری ماشین و یادگیری عمیق

محسوب می‌شود.

چارچوب روش خوشه‌بندی عمیق از سه عنصر تشکیل شده است: شبکه عصبی عمیق^۶، زیان شبکه^۷ و زیان خوشه‌بندی^۸ (شکل ۷-۲). در طراحی خوشه‌بندی عمیق از ساختار شبکه‌های عصبی عمیق برای

^۱ Classification

^۲ Constraint-based methods

^۳ Distance-based methods

^۴ Principal component analysis- PCA

^۵ Deep Clustering

^۶ Deep Neural Network- DNN

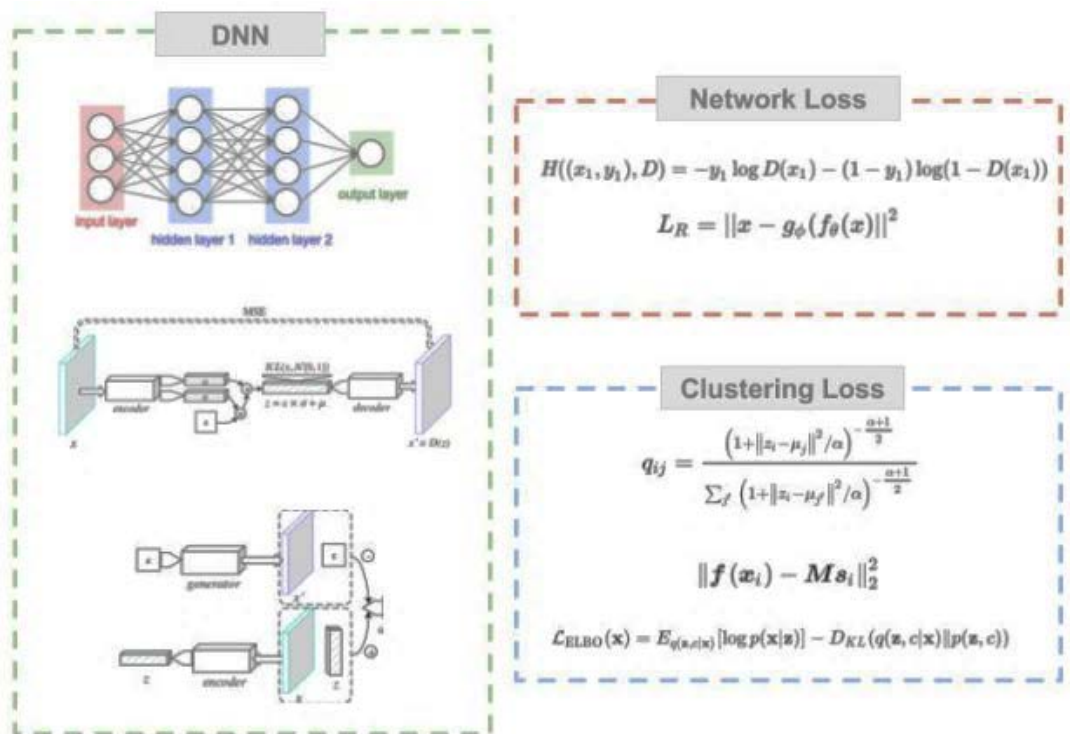
^۷ Network Loss

^۸ Clustering Loss

یادگیری ویژگی‌های داده‌ها استفاده می‌شود و با کمک تابع هدف (تابع زیان) زیر مسئله خوشه‌بندی حل خواهد شد (Aljalbout et al. 2019):

$$L(\theta) = \alpha L_c(\theta) + (1 - \alpha) L_n(\theta) \quad (13-2)$$

که در آن $L_c(\theta)$ تابع زیان خوشه‌بندی، $L_n(\theta)$ تابع زیان غیر خوشه‌بندی و $\alpha \in [0; 1]$ ضریب وزنی نقش این دو تابع زیان است که می‌تواند ثابت باشد و یا در مراحل مختلف الگوریتم تغییر کند (Aljalbout et al. 2019).



شکل ۷-۲- چارچوب روش خوشه‌بندی عمیق (Aljalbout et al. 2019)

از رویکردها و روش‌های متعددی برای خوشه‌بندی عمیق داده‌ها در ادبیات استفاده شده است که آنها را می‌توان از منظرهای مختلف دسته‌بندی کرد.

از منظر مدل‌سازی، روش‌های ارائه شده برای خوشه‌بندی عمیق در ادبیات را می‌توان در سه حوزه زیر دسته‌بندی کرد:

□ AutoEncoders based

- o Deep Embedded Clustering (Xie, Girshick, and Farhadi 2016)
- o Discriminately Boosted Clustering (F. Li, Qiao, and Zhang 2018)
- o Deep Clustering Network (B. Yang et al. 2017)

Generative Model Based □

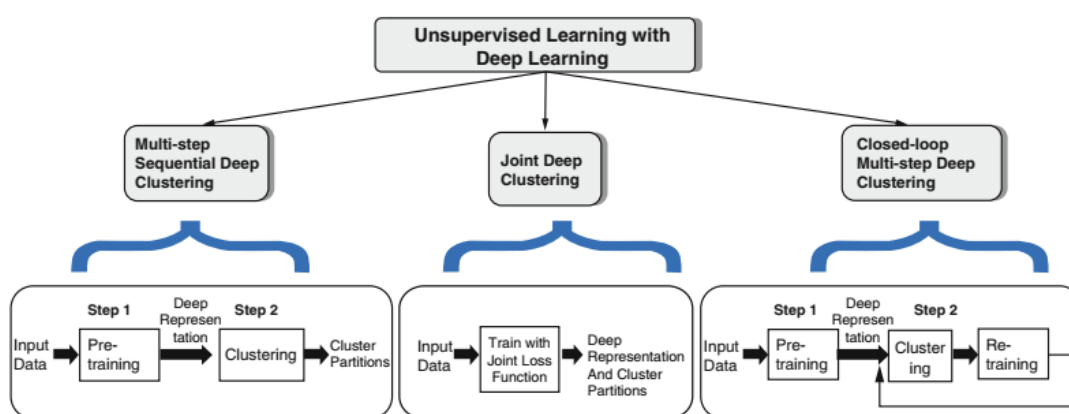
- Variational Deep Embedding (Jiang et al. 2016) ○
- Information Maximizing Generative Adversarial Network (Chen et al. 2016) ○

Direct Cluster Optimization □

- Joint Unsupervised Learning (JULE) (J. Yang, Parikh, and Batra 2016) ○
- Deep Adaptive Image Clustering (DAC) (Chang et al. 2017) ○
- Information Maximizing Self-Augmented Training (Hu et al. 2017) ○

از منظر گام‌های خوشه‌بندی و براساس پژوهش (Aljalbout et al. 2019) رویکردهای خوشه‌بندی عمیق را می‌توان در سه دسته دسته‌بندی کرد (شکل ۲-۸):

- رویکرد Sequential Multistep Deep Clustering: در این رویکرد خوشه‌بندی در دو گام انجام می‌شود. در گام نخست بازنمایی عمیق‌تری از داده ورودی محاسبه شده و سپس در گام دوم خوشه‌بندی بر روی این بازنمایی عمیق انجام می‌شود.
- رویکرد Joint Deep Clustering: در این رویکرد بازنمایی و خوشه‌بندی به صورت همزمان انجام می‌شود (به جای طی دو گام به صورت مجزا).
- رویکرد Closed-loop multistep Deep Clustering: این رویکرد مشابه رویکرد نخست است با این تفاوت که در رویکرد نخست هیچ بازگشتی وجود ندارد و گام‌ها رو به جلئی هستند اما در این رویکرد از نتایج خوشه‌بندی در باز-آموزش شبکه عمیق استفاده می‌شود.



شکل ۲-۸- رویکردهای خوشه‌بندی عمیق بر اساس پژوهش (Aljalbout et al., 2019)

از منظر ساختار شبکه در خوشه‌بندی عمیق از شبکه‌های زیر استفاده شده است (Min et al. 2018):

- Feedforward Fully-Connected Neural Network
- Feedforward Convolutional Neural Network
- Deep Belief Network
- Autoencoder
- GAN & VAE

آنچه در تمامی رویکردها و دسته‌بندی‌های اشاره شده مشترک است آن است که در تمامی این روش‌ها ابتدا بازنمایی از داده‌های ورودی انجام شده و سپس خوشه‌بندی بر روی این بازنمایی انجام می‌شود. چرا که ابعاد داده‌های بازنمایی شده عموماً از ابعاد داده‌های ورودی کمتر بوده و خوشه‌های پنهان در داده‌ها راحت‌تر کشف می‌شوند (Aljalbout et al. 2019).

۲-۵-۴: خلاصه‌سازی نتایج و داده‌ها

خلاصه‌سازی نتایج در خوشه‌بندی به معنای ارائه توضیحی با معنی و در عین حال مختصر برای هر خوشه است. در خوشه‌بندی متن این مساله را می‌توان به صورت اختصاص برچسب مناسب به اسناد هر دسته تعریف کرد. به عنوان نمونه بعد از دسته‌بندی اسناد مرتبط با عبارت «جگوار» بهتر است اسناد در دو دسته با نام‌های «حیوان» و «خودرو» برای کاربر نمایش داده شود. روش‌های مختلفی برای این منظور وجود دارد که بسته به کاربرد متفاوت هستند، به عنوان نمونه (Feldman and Sanger 2007):

- استفاده از عنوان اسنادی که مرکز خوشه هستند؛
- استفاده از مجموعه واژگان پر کاربرد در هر خوشه (۵ یا ۱۰ واژه پر کاربرد)؛
- تولید یک عبارت اسمی برای اعضای هر خوشه.

۲-۵-۵: ارزیابی

همه روش‌های خوشه‌بندی، فارغ از نوع، به دنبال بیشینه ساختن معیارهای زیر هستند (Zhai and Massung 2016):

- انسجام^۱: اعضای یک خوشه تا چه حد به هم شبیه هستند؟
- انفصال^۲: اعضای خوشه‌ها تا چه حد از هم نامتشابهند؟

^۱ Coherence

^۲ Separation

□ کاربردپذیری^۱: خوشه‌های بدست آمده تا چه حد برای کاربرد مورد نظر مفید هستند؟

براساس سه معیار فوق، معیارهای اندازه‌گیری متعددی توسعه داده شده‌اند که از آن میان می‌توان به معیارهای دقت (رابطه^{۱-۱})، صحت (رابطه^{۲-۱})، حساسیت (رابطه^{۳-۱})، تشخیص‌پذیری (رابطه^{۴-۱}) و ماتریس آشفتگی (شکل ۱-۴) اشاره کرد. همچنین اگر خوشه‌بندی با هدف تسهیل بازیابی اطلاعات استفاده شده باشد، معیارهایی مانند بازیابی^۲، F1 و یا MAP^۳ استفاده می‌شود (Zhai and Massung 2016).

۶-۲ خوشه‌بندی اسناد کوتاه

همانگونه که پیشتر اشاره شد، خوشه‌بندی اسناد متنی کوتاه با خوشه‌بندی اسناد معمول متفاوت است چرا که مسئله تنک بودن در این اسناد شایع‌تر و شدیدتر است؛ اغلب کلمات تنها یکبار در این اسناد تکرار می‌شوند و در نتیجه معیار TF-IDF کارایی لازم را نخواهد داشت. برای حل کردن این مسئله روش‌های مختلفی استفاده شده است؛ برخی تلاش کردند تا با کمک ویکی‌پدیا و یا آنتولوژی متن سند را توسعه داده یا بر غنای واژگان آن بیافزایند. این کار نیازمند برخورداری از دانش حوزه پردازش زبان طبیعی است و همچنین باعث افزایش پیچیدگی محاسباتی خواهد شد (xu et al. 2015). راه دیگر، بهبود و ارتقاء روش‌های معمول خوشه‌بندی متن است، مانند مدلسازی موضوعی احتمالی، الگوریتم‌های فراابتکاری و ... (Fuentealba, López, and Ponce 2021)، (Jia et al. 2018) و (S. Yang, Huang, and Cai 2019). برخی دیگر نیز از شبکه‌های عصبی، به خصوص شبکه‌های عمیق استفاده کرده‌اند به عنوان نمونه (Aljaber et al. 2010)، (Y. Kim 2014)، (xu et al. 2015)، (Song, He, and Fu 2015)، (P. Wang et al. 2015)، (P. Wang et al. 2016)، (Jinarat, (Shen et al. 2019) و Manaskasemsak, and Rungsawang 2018).

^۱ Utility

^۲ Recall

^۳ mean average precision- MAP

فصل سوم

روش پیشنهادی

۱-۳ مقدمه

همانگونه که در فصل نخست بیان شد، به منظور پیاده‌سازی روش خوشه‌بندی پیشنهادی لازم است گام‌های زیر انجام شوند:

- گردآوری داده‌های مورد نیاز
- پیشنهاد رویکرد خوشه‌بندی مناسب برای بکارگیری در سیستم پیشنهاددهنده متنی
- نگاشت ویژگی‌ها و معیارهای شباهت بر داده‌های موجود
- پیاده‌سازی رویکرد خوشه‌بندی پیشنهادی به صورت آزمایشی و ارزیابی

این فصل به بررسی سه گام نخست بالا اختصاص دارد و به گام نهایی، پیاده‌سازی رویکرد خوشه‌بندی پیشنهادی، در فصل بعد پرداخته خواهد شد.

۲-۳ داده‌های مورد پردازش

مجبی و همکاران (۱۴۰۰) در طرح پژوهشی خود مجموعه داده استاندارد فارسی را برای مساله استخراج کلیدواژه از مستندات علمی فارسی تهیه کردند. این مجموعه داده شامل اطلاعات مربوط به ۳۰۴۷ پایان‌نامه یا رساله فارسی است که از پایگاه اطلاعات علمی ایران، گنج، جمع‌آوری شده و مربوط به پایان‌نامه‌هایی هستند که طی سال‌های ۱۳۹۸-۱۴۰۰ در این پایگاه اضافه شده‌اند. ویژگی‌های اطلاعاتی این مجموعه داده به شرح زیر است:

توسعه رویکرد خوشه‌بندی برای پیشنهاد اسناد علمی فارسی ... □ ۳۸

- ☐ شناسه سند
 - ☐ عنوان
 - ☐ چکیده
 - ☐ موضوع (در دو سطح)
 - ☐ رشته
 - ☐ گرایش
 - ☐ کلیدواژه‌های نویسنده
 - ☐ کلیدواژه‌های نمایه‌ساز

همچنین، پراکندگی داده‌های این مجموعه داده در چهار حوزه مختلف فنی و مهندسی، علوم انسانی، علوم و علوم کاربردی، و هنر، براساس موضوع‌بندی موجود در پایگاه گنج، به شرح جدول زیر است:

جدول ۳-۱- تعداد اسناد در هر حوزه موضوعی

حوزه موضوعی	تعداد سند
علوم انسانی	۱۴۴۳
فنی و مهندسی	۹۵۹
علوم و علوم کاربردی	۴۵۷
هنر	۱۸۸

برشی از داده‌های این مجموعه در شکل ۳-۱ نمایش داده شده است.

[illegible]

شکل ۳-۱- برشی از مجموعه داده

از این مجموعه داده و با هدف خوشه‌بندی اسناد، دو ویژگی عنوان و چکیده انتخاب شدند و از دو ویژگی موضوع سطح اول و موضوع سطح دوم برای کنترل نتایج خوشه‌بندی استفاده شد.

بنابراین داده‌های در دست برای خوشه‌بندی اسناد فارسی شامل دو ویژگی عنوان و چکیده بوده و از دو ویژگی موضوع سطح اول و موضوع سطح دوم به عنوان برچسب داده‌ها استفاده شد.

۳-۳ رویکرد خوشه‌بندی پیشنهادی

همانگونه که در فصل دوم عنوان شد، برای خوشه‌بندی گام‌های زیر طی می‌شود (طی همه گام‌ها ضروری نیست) (Feldman and Sanger 2007):

- گام ۱: بازنمایی مسئله شامل استخراج و یا انتخاب ویژگی‌های داده‌ها
- گام ۲: انتخاب و تعریف معیار شباهت و نزدیکی در خوشه‌بندی
- گام ۳: خوشه‌بندی داده‌ها
- گام ۴: خلاصه‌سازی نتایج و داده‌ها
- گام ۵: ارزیابی

گام‌های طی شده برای خوشه‌بندی داده‌های مورد پردازش در ادامه تشریح می‌شوند:

۳-۳-۱ گام ۱: بازنمایی مسئله

همانگونه که در بخش ۱-۵-۲ اشاره شد، اسناد متنی را می‌توان یا به صورت داده‌های دودویی (وجود و یا عدم وجود واژه‌ای در متن) و یا به صورت متن (با سطح ریزدانگی حرف، واژه، جمله، پاراگراف، تک صفحه، بخش و یا کل سند) بازنمایی کرد (Aggarwal and Zhai 2013).

یکی از پرکاربردترین روش‌ها در حالت بازنمایی دودویی، بازنمایی TF-IDF است که به دلایل متعددی چون تنک بودن ماتریس داده‌ها و ابعاد بسیار زیاد آن، کمتر بودن مولفه‌های اساسی در داده‌ها به نسبت ویژگی‌ها و نیاز به نرمالسازی داده‌ها، نمی‌توان از این بازنمایی به صورت کارا برای خوشه‌بندی اسناد استفاده کرد. همچنین استفاده از روش‌های مبتنی بر کاهش ابعاد داده نیز به دلیل عدم در نظرگیری ارتباط بین واژگان و بهینه نبودن برای پردازش داده‌های حجیم نیز معقول به نظر نمی‌رسد. بعلاوه این که پیش‌پردازش داده‌های متنی فارسی به اندازه کافی زمان‌بر و پرچالش است.

با توجه به موارد عنوان شده، از رویکرد یادگیری ماشین و شبکه‌های عمیق برای بازنمایی اسناد موجود در مجموعه داده استفاده می‌شود. در ادبیات اغلب از رویکردهای word2vec، GloVe و BERT، با

توجه با کارایی بسیار خوب آنها، برای بازنمایی استفاده می‌شود. با این حال سطح ریزدانگی در این روش‌ها، «کلمه» است. این سطح ریزدانگی در هنگام پردازش یک متن و یا یک پاراگراف می‌تواند ناکارآمدی‌هایی را از منظر زمان پردازش و نتایج بوجود آورد.

برای این منظور لازم است از روش‌های بازنمایی با سطح ریزدانگی پاراگراف استفاده کرد. یکی از این روش‌ها، روش Doc2vec است که در ادامه روش Word2vec و توسط (Le and Mikolov 2014) توسعه داده شد. سطح ریزدانگی این روش جملات و پاراگراف است. علی‌رغم نتایج بسیار خوب Word2vec در بازنمایی کلمات، روش Doc2vec نتایج مطلوبی را در بازنمایی جملات خصوصاً در اسناد غیرانگلیسی زبان حاصل نکرد (Lau and Baldwin 2016).

SBERT^۱ یکی دیگر از روش‌های بازنمایی با سطح ریزدانگی پاراگراف است که برای بازنمایی متن در بیش از ۱۰۰ زبان و برای کاربردهای متعددی نظیر خوشه‌بندی، دسته‌بندی، مدلسازی موضوعی، جستجوی درون تصویر، جستجوی معنایی و جستجوی بین زبانی توسعه یافته است. همچنین تمامی کدها و مجموعه داده‌های مورد استفاده در این روش در پایگاه <https://sbert.net> در دسترس قرار دارد (Reimers and Gurevych 2020).

بر اساس موارد مطرح شده، در این پژوهش از روش SBERT برای بازنمایی داده‌ها استفاده می‌شود.

۳-۳-۲ گام ۲: معیار شباهت

معیار شباهت معیاری برای سنجش میزان شباهت اجزای درون یک خوشه بوده و بایستی بر اساس ویژگی‌های داده‌ها و خصوصیات روش خوشه‌بندی انتخاب شود. معیار فاصله یکی از معیارهای اصلی برای محاسبه شباهت بین دو داده است و انواع مختلفی از معیار فاصله در ادبیات توسعه داده شده است که در بخش ۲-۵-۲ به آنها اشاره شد.

از آنجا که داده‌های حاصل از بازنمایی، داده‌های عددی هستند از معیارهای فاصله متنوعی می‌توان استفاده کرد. اما نکته قابل توجه و مهم آن است که خروجی بازنمایی، ماتریسی با ابعاد بسیار بزرگ (بیشتر از ۲۰۰ بعد) خواهد بود و در نتیجه بایستی در انتخاب نوع فاصله دقت کرد. نفرین بعد^۲ موضوعی است که در انتخاب نوع فاصله باید بدان دقت کرد چرا که بسیاری از روش‌های خوشه‌بندی کارایی خود را در مواجهه با داده‌هایی با ابعاد زیاد از دست داده و از توانایی آنها در تفکیک داده‌ها کاسته

^۱ Sentence-BERT

^۲ The curse of dimensionality

می‌شود. (Aggarwal and Zhai 2013) نشان دادند که در حالت استفاده از فاصله مینکویسکی (رابطه (۱-۲)) با افزایش ابعاد داده، بکارگیری $p \in (0,1)$ می‌تواند کاراتر باشد.

برای حل این مشکل می‌توان از توابع فاصله‌ای دیگری استفاده کرد که کمتر تحت تاثیر ابعاد داده‌ها هستند، نظیر فاصله کسینوسی (رابطه (۲-۹)) یا آنتروپی نسبی (رابطه (۲-۱۲)).

۳-۳-۳: خوشه‌بندی داده‌ها

در بخش‌های قبلی به تفصیل به روش‌های پرکاربرد در خوشه‌بندی اشاره شد. به دلیل ابعاد زیاد داده‌های بازنمایی شده، روش خوشه‌بندی عمیق کارایی بهتری نسبت به سایر روش‌های خوشه‌بندی دارد. لذا در این پژوهش از روش خوشه‌بندی عمیق به صورت Sequential Multistep Deep Clustering (بخش ۳-۵-۲-) استفاده شد.

در این روش، ابتدا داده‌های ورودی با کمک روش SBERT بازنمایی شده و سپس با استفاده از الگوریتم‌های K-means (جدول ۲-۳)، FCM (رابطه (۳-۱)) و REFCM (Zarinbal, Fazel و Zarandi, and Turksen 2014) (رابطه (۳-۴)) خوشه‌بندی می‌شوند. الگوریتم نخست (K-means)، روشی قطعی برای خوشه‌بندی است در صورتی که در دو الگوریتم بعدی (FCM و REFCM) از منطق فازی برای خوشه‌بندی استفاده شده است.

جدول ۲-۳- شبه کد الگوریتم k-means

k-means
تعیین کنید:
تعداد خوشه‌ها، k
مراکز خوشه‌ها، $c_1, c_2, \dots, c_k$
$t = 1$
گام‌های زیر را انجام دهید تا وقتی $\ c_j^t - c_j^{t-1}\ \leq \varepsilon \quad \forall j \in [1, k]$
گام ۱: فاصله تمام داده‌ها را تا تمامی مراکز خوشه محاسبه کنید، $d_{ik}^2, i: 1 \dots N$
گام ۲: داده‌ها را به نزدیک‌ترین مرکز خوشه تخصیص دهید
گام ۳: مراکز خوشه را محاسبه کنید
گام ۴: $t = t + 1$

مدل ریاضی الگوریتم Fuzzy c-means (FCM) به شرح زیر است:

$$\min \sum_{i=1}^c \sum_{j=1}^N u_{ij}^m d_{ij}^2$$

$$s.t. \begin{cases} \sum_{i=1}^c u_{ij} = 1 & \forall j \\ u_{ij} \in \{0,1\} & \forall j, i \end{cases} \quad (1-3)$$

که در آن m درجه فازی و u_{ij} تابع عضویت داده j ام در خوشه i ام بوده و از رابطه (۲-۳) حاصل می‌شود. مراکز خوشه‌ها نیز با کمک رابطه (۳-۳) بروزرسانی می‌شوند.

$$u_{ij} = \frac{1}{\sum_{k=1}^c \left(\frac{\|x_j - c_i\|}{\|x_j - c_k\|} \right)^{\frac{2}{m-1}}} \quad (2-3)$$

$$c_i = \frac{\sum_j u_{ij}^m x_j}{\sum_j u_{ij}^m} \quad (3-3)$$

مدل ریاضی الگوریتم Relative entropy Fuzzy c-means (REFCM) نیز به شرح زیر است (Zarinbal, Fazel Zarandi, and Turksen 2014):

$$\min J(U, V, c) = \sum_{i=1}^c \sum_{j=1}^N u_{ij}^m d_{ij}^2 - \theta \left(\sum_{j=1}^N \sum_{i=1}^c \sum_{\substack{k=1 \\ k \neq i}}^c u_{ij} \ln \left(\frac{u_{ij}}{u_{kj}} \right) \right)$$

$$s.t. \begin{cases} \sum_{i=1}^c u_{ij} = 1 & \forall j \\ 0 < \sum_{j=1}^N u_{ij} & \forall i \\ u_{ij} \in [0,1] & \forall i, j \end{cases} \quad (4-3)$$

که در آن m درجه فازی و u_{ij} تابع عضویت داده j ام در خوشه i ام بوده و از رابطه (۵-۳) حاصل می‌شود. مراکز خوشه‌ها نیز با کمک رابطه (۳-۳) بروزرسانی می‌شوند.

$$u_{ij} = \left(\frac{\frac{-m(m-1)d_{ij}^2}{\theta}}{W_0 \left[\frac{-m(m-1)d_{ij}^2}{\theta} \exp \left(-(m-1) \left(\frac{\lambda_j - \theta \sum_{\substack{k=1 \\ k \neq i}}^c \ln(u_{kj}) + \theta \right)}{\theta} \right) \right]} \right)^{\frac{1}{m-1}} \quad (5-3)$$

$W_0(\cdot)$ تابع لامبرت^۱ است با سه مشخصه $W_0(0) = 0$ ، $W_0(\cdot) \geq -1$ و $W_0(-\frac{1}{e}) = -1$. همچنین در همه روابط بالا d_{ij} بیانگر فاصله داده زام از مرکز خوشه i ام، c بیانگر تعداد خوشه‌ها و N تعداد داده‌هاست.

۳-۳-۴ گام ۴: ارزیابی

همانطور که در بخش ۵-۵-۲ اشاره شده، برای ارزیابی نتایج خوشه‌بندی عموماً از معیارهای دقت (رابطه ۱-۱)، صحت (رابطه ۲-۱)، حساسیت (رابطه ۳-۱)، تشخیص‌پذیری (رابطه ۴-۱) و ماتریس آشفتگی (شکل ۴-۱) استفاده می‌شود. اما برای داده‌های بزرگ و روش یادگیری عمیق، عموماً از معیار اصلاح شده دقت خوشه‌بندی بدون نظارت^۲ (ACC) (رابطه ۶-۳) استفاده می‌شود.

$$ACC = \max_m \frac{\sum_{i=1}^n 1\{y_i = f(c_i)\}}{n} \quad (6-3)$$

که در آن $f(\bullet)$ تابع نگاشت برچسب خوشه بدست آمده از خوشه‌بندی به برچسب از پیش داشته y_i برای هر خوشه است.

در فصل بعد، گام‌های ذکر شده بر داده‌های مورد پردازش اعمال شده و نتایج ارائه خواهند شد.

^۱ Lambert-W function

^۲ Unsupervised Clustering Accuracy (ACC)

فصل چهارم

پیاده‌سازی رویکرد خوشه‌بندی پیشنهادی

۱-۴ مقدمه

این فصل به پیاده‌سازی رویکرد خوشه‌بندی پیشنهادی به صورت آزمایشی و ارزیابی نتایج آن اختصاص دارد. همانطور که در فصل قبل بیان شد برای پیاده‌سازی این رویکرد از اطلاعات مربوط به ۳۰۴۷ پایان‌نامه یا رساله فارسی که از پایگاه اطلاعات علمی ایران، گنج طی سال‌های ۱۳۹۸-۱۴۰۰ جمع‌آوری شده استفاده شده است. گام‌های طی شده در ادامه تشریح خواهند شد.

۲-۴ گام ۱: بازنمایی مسئله

طبق آنچه در فصل قبل بیان شد، از روش SBERT برای بازنمایی داده‌ها استفاده می‌شود. سطح ریزدانگی این روش پاراگراف است و برای بازنمایی متن در بیش از ۱۰۰ زبان و برای کاربردهای متعددی نظیر خوشه‌بندی، دسته‌بندی، مدلسازی موضوعی، جستجوی درون تصویر، جستجوی معنایی و جستجوی بین زبانی توسعه یافته است.

در این پژوهش و به منظور بازنمایی داده‌ها از ساختار زیر استفاده شده است:

```
sentence_transformers.SentenceTransformer (paraphrase-multilingual-  
mpnet-base-v2, device = 'cuda')
```

خروجی این بازنمایی ماتریسی متشکل از ۳۰۴۷ سطر و ۷۶۸ ستون است. به بیانی دیگر، روش SBERT با کمک مدل از قبل آموزش دیده خود، توانست داده‌های ورودی را با ۷۶۸ ویژگی بازنمایی کند. بخشی از این بازنمایی در شکل ۴-۱ نمایش داده شده است.

	A	B	C	D	E	F	G
1		0	1	2	3	4	5
2	0	tensor(0.1533, device='cuda:0')	tensor(-0.0934, device='cuda:0')	tensor(-0.0076, device='cuda:0')	tensor(-0.0054, device='cuda:0')	tensor(-0.0977, device='cuda:0')	tensor(-0.0363, device='cuda:0')
3	1	tensor(0.0714, device='cuda:0')	tensor(-0.0904, device='cuda:0')	tensor(-0.0134, device='cuda:0')	tensor(-0.0643, device='cuda:0')	tensor(-0.0975, device='cuda:0')	tensor(-0.0717, device='cuda:0')
4	2	tensor(-0.0411, device='cuda:0')	tensor(0.1793, device='cuda:0')	tensor(-0.0122, device='cuda:0')	tensor(-0.0953, device='cuda:0')	tensor(0.0725, device='cuda:0')	tensor(-0.0592, device='cuda:0')
5	3	tensor(0.0061, device='cuda:0')	tensor(-0.0353, device='cuda:0')	tensor(-0.0095, device='cuda:0')	tensor(-0.0024, device='cuda:0')	tensor(0.0936, device='cuda:0')	tensor(-0.0264, device='cuda:0')
6	4	tensor(0.0245, device='cuda:0')	tensor(0.0291, device='cuda:0')	tensor(-0.0109, device='cuda:0')	tensor(0.0386, device='cuda:0')	tensor(-0.0180, device='cuda:0')	tensor(-0.0442, device='cuda:0')
7	5	tensor(0.0052, device='cuda:0')	tensor(0.0203, device='cuda:0')	tensor(-0.0126, device='cuda:0')	tensor(-0.0063, device='cuda:0')	tensor(0.0484, device='cuda:0')	tensor(-0.0173, device='cuda:0')
8	6	tensor(-0.0146, device='cuda:0')	tensor(-0.0416, device='cuda:0')	tensor(-0.0131, device='cuda:0')	tensor(-0.0003, device='cuda:0')	tensor(-0.0693, device='cuda:0')	tensor(-0.0176, device='cuda:0')
9	7	tensor(-0.0681, device='cuda:0')	tensor(-0.0897, device='cuda:0')	tensor(-0.0046, device='cuda:0')	tensor(0.0882, device='cuda:0')	tensor(-0.1074, device='cuda:0')	tensor(-0.0122, device='cuda:0')
10	8	tensor(0.0311, device='cuda:0')	tensor(0.0963, device='cuda:0')	tensor(-0.0105, device='cuda:0')	tensor(0.0500, device='cuda:0')	tensor(-0.0964, device='cuda:0')	tensor(0.0289, device='cuda:0')
11	9	tensor(0.0232, device='cuda:0')	tensor(-0.0127, device='cuda:0')	tensor(-0.0028, device='cuda:0')	tensor(-0.0505, device='cuda:0')	tensor(0.0585, device='cuda:0')	tensor(-0.0053, device='cuda:0')
12	10	tensor(-0.0703, device='cuda:0')	tensor(0.0674, device='cuda:0')	tensor(-0.0123, device='cuda:0')	tensor(-0.0010, device='cuda:0')	tensor(-0.0423, device='cuda:0')	tensor(0.0860, device='cuda:0')
13	11	tensor(-0.0604, device='cuda:0')	tensor(0.1129, device='cuda:0')	tensor(-0.0079, device='cuda:0')	tensor(-0.0867, device='cuda:0')	tensor(0.0236, device='cuda:0')	tensor(-0.0365, device='cuda:0')
14	12	tensor(-0.0128, device='cuda:0')	tensor(-0.0280, device='cuda:0')	tensor(-0.0106, device='cuda:0')	tensor(0.0220, device='cuda:0')	tensor(0.0799, device='cuda:0')	tensor(0.1130, device='cuda:0')
15	13	tensor(0.0632, device='cuda:0')	tensor(0.0107, device='cuda:0')	tensor(-0.0190, device='cuda:0')	tensor(0.0538, device='cuda:0')	tensor(-0.0210, device='cuda:0')	tensor(-0.0146, device='cuda:0')
16	14	tensor(0.0652, device='cuda:0')	tensor(-0.0144, device='cuda:0')	tensor(-0.0080, device='cuda:0')	tensor(-0.0078, device='cuda:0')	tensor(-0.0778, device='cuda:0')	tensor(-0.0279, device='cuda:0')
17	15	tensor(-0.0126, device='cuda:0')	tensor(-0.0864, device='cuda:0')	tensor(-0.0101, device='cuda:0')	tensor(-0.0322, device='cuda:0')	tensor(0.0114, device='cuda:0')	tensor(0.0282, device='cuda:0')
18	16	tensor(-0.0088, device='cuda:0')	tensor(-0.1825, device='cuda:0')	tensor(-0.0117, device='cuda:0')	tensor(-0.0660, device='cuda:0')	tensor(0.0916, device='cuda:0')	tensor(-0.0013, device='cuda:0')
19	17	tensor(-0.0126, device='cuda:0')	tensor(-0.0021, device='cuda:0')	tensor(-0.0114, device='cuda:0')	tensor(-0.0242, device='cuda:0')	tensor(-0.0077, device='cuda:0')	tensor(0.0282, device='cuda:0')
20	18	tensor(-0.0197, device='cuda:0')	tensor(0.0265, device='cuda:0')	tensor(-0.0085, device='cuda:0')	tensor(-0.1133, device='cuda:0')	tensor(0.0798, device='cuda:0')	tensor(0.0227, device='cuda:0')
21	19	tensor(0.0108, device='cuda:0')	tensor(0.1366, device='cuda:0')	tensor(-0.0135, device='cuda:0')	tensor(-0.0188, device='cuda:0')	tensor(-0.0529, device='cuda:0')	tensor(0.0051, device='cuda:0')
22	20	tensor(0.0525, device='cuda:0')	tensor(-0.0243, device='cuda:0')	tensor(-0.0149, device='cuda:0')	tensor(0.0497, device='cuda:0')	tensor(0.0096, device='cuda:0')	tensor(0.0289, device='cuda:0')
23	21	tensor(0.0784, device='cuda:0')	tensor(0.1310, device='cuda:0')	tensor(-0.0149, device='cuda:0')	tensor(0.0042, device='cuda:0')	tensor(-0.0941, device='cuda:0')	tensor(0.0624, device='cuda:0')

شکل ۴-۱- بخشی از بازنمایی داده‌های ورودی با کمک روش SBERT

۴-۳ گام ۲: معیار شباهت

معیار شباهت معیاری برای سنجش میزان شباهت اجزای درون یک خوشه بوده و بایستی بر اساس ویژگی‌های داده‌ها و خصوصیات روش خوشه‌بندی انتخاب شود. از آنجا که داده‌های بازنمایی شده از ابعاد بزرگی برخوردار هستند (۷۶۸ بعد)، علاوه بر استفاده از فاصله مینکویسکی در حالت $p \in (0,1)$ می‌توان از توابع دیگری نظیر فاصله کسینوسی (رابطه (۲-۹)) یا آنتروپی نسبی (رابطه (۲-۱۲)) استفاده کرد. لذا لازم است در مدلسازی روش خوشه‌بندی دقت لازم صورت گیرد.

۴-۴ گام ۳: خوشه‌بندی داده‌ها

به منظور خوشه‌بندی داده‌ها و مقایسه روش‌های موجود، سه الگوریتم خوشه‌بندی K-means (جدول ۳-۲)، FCM (رابطه (۳-۱)) و REFCM (رابطه (۳-۴)) در نرم‌افزار Python پیاده‌سازی شد و نتایج بدست آمد. همچنین از آنجا که ممکن بود ابعاد زیاد داده‌ها نتایج را تحت تاثیر قرار دهد، الگوریتم FCM با کمک دو فاصله اقلیدسی و کسینوسی پیاده‌سازی شد.

یکی از نکات مهم در خوشه‌بندی، تعیین تعداد خوشه‌هاست که با توجه به تعداد بالای داده‌ها و عدم دانش پیشین، باید به صورت تجربی تعیین شود. یکی از راهکارهای مورد استفاده در این رابطه استفاده از شاخص سیلوئت^۱ است (Handl and Knowles 2006). در این روش، داده‌ها را با تعداد مختلفی از خوشه‌ها خوشه‌بندی کرده و تعداد خوشه‌ای که بیشترین مقدار را برای این شاخص فراهم کند، مناسب‌ترین تعداد خوشه‌است.

^۱ Silhouette

با تکرار خوشه‌بندی بر روی داده‌های موجود با تعداد خوشه‌های مختلف، شاخص سیلوئت بهترین مقدار را در $c = 13$ و $c = 11$ حاصل کرد.

در ادامه و با کمک این دو عدد ۳ الگوریتم خوشه‌بندی اجرا شد و نتایج بدست آمد.

۵-۴ گام ۴: ارزیابی

همانگونه که ذکر شد، از معیارهای دقت، صحت، حساسیت و تشخیص‌پذیری برای ارزیابی نتایج خوشه‌بندی استفاده می‌شود اما از آنجا که در داده‌های بزرگ، آگاهی از برچسب‌های همه داده‌ها ممکن نیست و یا برچسب‌های کلی برای داده‌ها وجود دارد معیار دقت اصلاح شده ACC (رابطه (۳-۶)) استفاده می‌شود. جدول ۴-۱ و جدول ۴-۲ نتایج بدست آمده را برای $c = 13$ و $c = 11$ گزارش می‌کنند.

جدول ۴-۱- نتایج بدست آمده برای $c=13$

REFCM	FCM-cosine	FCM-euclidean	k-means		
78%	92%	88%	85%	1	Accuracy with-in Class
70%	86%	86%	86%	2	
75%	87%	83%	83%	3	
75%	70%	82%	89%	4	
51%	56%	91%	93%	5	
92%	86%	86%	86%	6	
66%	88%	86%	89%	7	
91%	92%	92%	93%	8	
50%	41%	70%	77%	9	
61%	61%	75%	72%	10	
74%	84%	80%	82%	11	
96%	86%	92%	92%	12	
100%	92%	88%	87%	13	
78%	79%	84%	85%	Total ACC	

جدول ۴-۲- نتایج بدست آمده برای $c=11$

REFCM	FCM-cosine	FCM-euclidean	k-means		
82%	82%	94%	93%	1	accuracy with-in Class
95%	95%	89%	89%	2	
75%	75%	68%	68%	3	
90%	90%	93%	93%	4	
95%	95%	93%	93%	5	
88%	88%	79%	79%	6	
93%	93%	93%	93%	7	
67%	67%	75%	75%	8	
96%	96%	94%	4%	9	
88%	88%	87%	87%	10	
96%	96%	88%	88%	11	
86%	86%	85%	80%	Total ACC	

براساس نتایج بدست آمده می‌توان گفت که تمامی الگوریتم‌های مورد استفاده نتایج خوبی را حاصل کردند و در حالت $c=11$ و با استفاده از FCM (فاصله کسینوسی) و REFCM نتایج بهتری نسبت به $c=13$ حاصل شد. تفسیر خوشه‌های بدست آمده در حالت $c=11$ به شرح جدول ۴-۳ است.

جدول ۴-۳- تفسیر خوشه‌های بدست آمده در حالت $c=11$

موضوع					خوشه
برق					1 ریاضی رباتیک
عمران					2 سازه
کشاورزی	محیط زیست	انرژی	ژنتیک	بیوشیمی	3 پزشکی
نقاشی					4 هنر معماری
رسانه					5 تاریخ فقه ادبیات
روابط بین الملل					6 حقوق فلسفه معارف منطق
					7 روانشناسی آموزش
فناوری اطلاعات					8 کسب و کار بازاریابی مدیریت دانش سازمانی
گردشگری					9 مسائل اجتماعی شهرنشینی
					10 اقتصاد بیمه
پلاستیک					11 مواد آلی و پلیمر نانو نور



هدف اصلی در این طرح، توسعه و بکارگیری رویکرد خوشه‌بندی برای پیشنهاد اسناد علمی فارسی به کاربر بود. بدین منظور ابتدا ادبیات رویکردهای خوشه‌بندی متن در سیستم‌های پیشنهاددهنده و ویژگی‌های مورد استفاده در هریک از رویکردها بررسی شده و رویکرد پیشنهادی معرفی شد. سپس با استفاده از فراداده‌های پایان‌نامه‌ها و رساله‌های فارسی پایگاه گنج، شامل اطلاعات مربوط به ۳۰۴۷ پایان‌نامه یا رساله فارسی رویکرد پیشنهادی پیاده‌سازی و ارزیابی شد.

بر اساس ادبیات مربوط به خوشه‌بندی داده‌های بزرگ، سه معیار شباهت مینکویسکی، کسینوسی و آنتروپی نسبی به عنوان معیارهای فاصله یا شباهت انتخاب شدند. با کمک این معیارهای شباهت سه الگوریتم خوشه‌بندی K-means، FCM و REFCM در نرم‌افزار Python پیاده‌سازی شده و نتایج بدست آمد. براساس نتایج بدست آمده می‌توان گفت که تمامی الگوریتم‌های مورد استفاده نتایج خوبی را حاصل کردند و در حالت $c=11$ و با استفاده از FCM (فاصله کسینوسی) و REFCM نتایج بهتری نسبت به $c=13$ حاصل شد. از آنجا که در این حالت هزینه محاسبات به مراتب کمتر $c=13$ خواهد بود از تفسیر خوشه‌های بدست آمده در حالت $c=11$ به عنوان مفسر داده‌های مورد پردازش استفاده شد.

بر اساس تفسیر بدست آمده می‌توان پایان‌نامه یا رساله‌های حوزه‌های زیر را در یک خوشه مشترک قرار داد و در صورتی که کاربر جستجویی را در یکی از زیرحوزه‌های زیر انجام داده باشد، نتایج مرتبط را بر اساس دسته‌بندی زیر به او نمایش داد:

- ریاضی، برق، رباتیک
- سازه، عمران
- پزشکی، بیوشیمی، ژنتیک، انرژی، محیط زیست
- هنر، معماری، نقاشی
- تاریخ، فقه، ادبیات، رسانه
- حقوق، فلسفه، معارف، منطق، روابط بین‌الملل
- روانشناسی، آموزش
- کسب و کار، بازاریابی، مدیریت، دانش سازمانی، فناوری اطلاعات
- مسائل اجتماعی، شهرنشینی، گردشگری
- اقتصاد، بیمه
- مواد آلی و پلیمر، نانو، نور، پلاسما

پیشنهاد پژوهشی و کاربردی

همانگونه که در فصل اول بیان شد، سیستم‌های پیشنهاددهنده پیشنهادهایی متناسب با نیاز کاربر به او ارائه داده و به تصمیم‌گیری برای انتخاب موارد مورد نیاز یا مورد علاقه کاربر کمک می‌کنند. تحلیل سند منتخب کاربر و استفاده از ویژگی‌های خاص آن برای پیشنهاد اسناد مشابه یکی از روش‌ها است.

براین اساس و در ادامه این پژوهش می‌توان به موضوع ساخت و توسعه سیستم پیشنهاددهنده‌ای براساس سیستم‌های پیشنهاددهنده مبتنی بر محتوی پرداخت که در آنها، ویژگی‌های سند(های) منتخب کاربر فعلی بررسی شده و مواردی به او پیشنهاد شود که با این ویژگی‌ها همخوانی یا شباهت داشته باشند.

زرین‌بال، مرضیه. ۱۳۹۸. بررسی موتور جست‌وجوی گنج به منظور بهبود بازیابی اطلاعات در زبان فارسی، پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک).

مجیبی، آزاده، عمار جلالی‌منش. ۱۳۹۷. ارائه روشی هوشمند برای استخراج کلیدواژه از مستندات علمی زبان فارسی براساس سیستم‌های پیشنهاددهنده، پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک).

مجیبی، آزاده، عمار جلالی‌منش، مرضیه زرین‌بال، زهرا دهسرای، علی سلیمی‌صدر، حمید حسنی. ۱۴۰۰. تهیه مجموعه داده استاندارد فارسی برای مساله استخراج خودکار کلیدواژه از اسناد علمی، پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک).

هوشینو، <https://hooshio.com/>. آنچه-در-مورد-مدل-زبانی-bert-نمی-دانیم / (دسترسی در ۱۴۰۰/۱۱/۲۷)

Aggarwal, Charu C., and ChengXiang Zhai, eds. 2013. *Mining Text Data*. Springer.

Aljaber, Bader, Nicola Stokes, James Bailey, and Jian Pei. 2010. "Document Clustering of Scientific Texts Using Citation Contexts." *Information Retrieval* 13(2): 101–31.

Aljalbout, Elie et al. 2019. "Clustering with Deep Learning: Taxonomy and New Methods." In *Clustering Methods for Big Data Analytics*, Springer Nature Switzerland, 73–89. <http://arxiv.org/abs/1801.07648>.

Andritsos, Periklis, Panayiotis Tsaparas, Renée J Miller, and Kenneth C Sevcik. 2004. "LIMBO: Scalable Clustering of Categorical Data." In *International Conference on Extending Database Technology*, Springer, 123–46.

Burke, Robin. 2007. "Hybrid Web Recommender Systems." In *The Adaptive Web, Methods and Strategies of Web Personalization*, eds. Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl. Berlin: Springer, 377–408.

Chang, Jianlong et al. 2017. "Deep Adaptive Image Clustering." In *Proceedings of the IEEE International Conference on Computer Vision*, , 5879–87.

Chen, Xi et al. 2016. "Infogan: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets." In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, , 2180–88.

Devlin, Jacob, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 1*: 4171–86.

- FarsiYar. 2020. "Text Mining." <https://blog.text-mining.ir/text-mining-process-tapz6y4qbz8r/> (June 13, 2021).
- Feldman, Ronen, and James Sanger. 2007. *The Text Mining Handbook*. Cambridge University Press.
- Fuentealba, Diego, Mario López, and Héctor Ponce. 2021. "Effects on Time and Quality of Short Text Clustering during Real-Time Presentations." *IEEE Latin America Transactions* 19(8): 1391–99.
- Gibson, David, Jon Kleinberg, and Prabhakar Raghavan. 2000. "Clustering Categorical Data: An Approach Based on Dynamical Systems." *The VLDB Journal* 8(3): 222–36.
- Grainger, Trey, and Timothy Potter. 2014. *Solr in Action*. New York: Manning Publications Co. <https://beta.manning.com/dashboard/%5Cnpapers2://publication/uuid/17DA7ECA-A9EC-4DE8-924B-5016C8859C40>.
- Guha, Sudipto, Rajeev Rastogi, and Kyuseok Shim. 2001. "Cure: An Efficient Clustering Algorithm for Large Databases." *Information systems* 26(1): 35–58.
- Handl, Julia, and Joshua Knowles. 2006. "Multi-Objective Clustering and Cluster Validation." In *Multi-Objective Machine Learning*, Springer, 21–47.
- Hu, Weihua et al. 2017. "Learning Discrete Representations via Information Maximizing Self-Augmented Training." In *International Conference on Machine Learning*, PMLR, 1558–67.
- Huang, Anna. 2008. "Similarity Measures for Text Document Clustering." In *Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008)*, Christchurch, New Zealand, 49–56.
- Jia, Caiyan, Matthew B. Carson, Xiaoyang Wang, and Jian Yu. 2018. "Concept Decompositions for Short Text Clustering by Identifying Word Communities." *Pattern Recognition* 76: 691–703.
- Jiang, Zhuxi et al. 2016. "Variational Deep Embedding: An Unsupervised and Generative Approach to Clustering." *arXiv preprint arXiv:1611.05148*.
- Jinarat, Supakpong, Bundit Manaskasemsak, and Arnon Rungsawang. 2018. "Short Text Clustering Based on Word Semantic Graph with Word Embedding Model." In *2018 Joint 10th International Conference on Soft Computing and Intelligent Systems and 19th International Symposium on Advanced Intelligent Systems, SCIS-ISIS 2018*, IEEE, 1427–32.
- Jomsri, Pijitra. 2018. "FUCL Mining Technique for Book Recommender System in Library Service." *Procedia Manufacturing* 22: 550–57. <http://www.sciencedirect.com/science/article/pii/S2351978918303779>.
- Kim, Donghyun, Chanyoung Park, Jinoh Oh, and Hwanjo Yu. 2017. "Deep Hybrid Recommender Systems via Exploiting Document Context and Statistics of

- Items.” *Information Sciences* 417: 72–87.
<http://www.sciencedirect.com/science/article/pii/S0020025516318539>.
- Kim, Yoon. 2014. “Convolutional Neural Networks for Sentence Classification.” In *Conference on Empirical Methods in Natural Language Processing*, , 1746–51.
- Lau, Jey Han, and Timothy Baldwin. 2016. “An Empirical Evaluation of Doc2vec with Practical Insights into Document Embedding Generation.” In *1st Workshop on Representation Learning for NLP*, , 78–86.
- Le, Quoc, and Tomas Mikolov. 2014. “Distributed Representations of Sentences and Documents.” In *Proceedings of the 31st International Conference on Machine Learning*, Beijing, China., 9.
- Li, Fengfu, Hong Qiao, and Bo Zhang. 2018. “Discriminatively Boosted Image Clustering with Fully Convolutional Auto-Encoders.” *Pattern Recognition* 83: 161–73.
- Li, Ming, Ying Li, Wangqin Lou, and Lisheng Chen. 2020. “A Hybrid Recommendation System for Q&A Documents.” *Expert Systems with Applications* 144: 113088.
<http://www.sciencedirect.com/science/article/pii/S095741741930805X>.
- Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. “Efficient Estimation of Word Representations in Vector Space.” *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*: 1–12.
- Min, Erxue et al. 2018. “A Survey of Clustering with Deep Learning: From the Perspective of Network Architecture.” *IEEE Access* 6: 39501–14.
- Nailli, Marwa, Anja Habacha Chaibi, and Henda Hajjami Ben Ghezala. 2017. “Comparative Study of Word Embedding Methods in Topic Segmentation.” *Procedia Computer Science* 112: 340–49.
<http://dx.doi.org/10.1016/j.procs.2017.08.009>.
- De Nart, D, and C Tasso. 2014. “A Personalized Concept-Driven Recommender System for Scientific Libraries.” *Procedia Computer Science* 38: 84–91.
<http://www.sciencedirect.com/science/article/pii/S1877050914013751>.
- Park, Deuk Hee, Hyea Kyeong Kim, Il Young Choi, and Jae Kyeong Kim. 2012. “A Literature Review and Classification of Recommender Systems Research.” *Expert systems with applications* 39(11): 10059–72.
- Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. “GloVe: Global Vectors for Word Representation.” In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, , 1532–43.
- Peters, Matthew E., Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. “Semi-Supervised Sequence Tagging with Bidirectional Language Models.” In

- 55th Annual Meeting of the Association for Computational Linguistics, , 1756–65.
- Pradhan, Tribikram, and Sukomal Pal. 2019. “CNAVER: A Content and Network-Based Academic VEnue Recommender System.” *Knowledge-Based Systems*: 105092. <http://www.sciencedirect.com/science/article/pii/S0950705119304691>.
- Reimers, Nils, and Iryna Gurevych. 2020. “Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks.” In *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, , 3982–92.
- Ricci, Francesco, Lior Rokach, Bracha Shapira, and Paul B. Kantor. 2011. *Recommender Systems Handbook*. New York, NY, USA: Springer.
- Saleh, Ahmed I, Ali I El Desouky, and Shereen H Ali. 2015. “Promoting the Performance of Vertical Recommendation Systems by Applying New Classification Techniques.” *Knowledge-Based Systems* 75: 192–223. <http://www.sciencedirect.com/science/article/pii/S0950705114004328>.
- Salehi, Mojtaba, Mohammad Pourzaferani, and Seyed Amir Razavi. 2013. “Hybrid Attribute-Based Recommender System for Learning Material Using Genetic Algorithm and a Multidimensional Information Model.” *Egyptian Informatics Journal* 14(1): 67–78. <http://www.sciencedirect.com/science/article/pii/S1110866512000527>.
- Saxena, Amit et al. 2017. “A Review of Clustering Techniques and Developments.” *Neurocomputing* 267: 664–81. <http://dx.doi.org/10.1016/j.neucom.2017.06.053>.
- Shen, Ying et al. 2019. 514 Lecture Notes in Electrical Engineering *Improving Medical Short Text Classification with Semantic Expansion Using Word-Cluster Embedding*. Springer Singapore. http://dx.doi.org/10.1007/978-981-13-1056-0_41.
- Song, Jiaying, Yu He, and Guohong Fu. 2015. “Polarity Classification of Short Product Reviews via Multiple Cluster-Based SVM Classifiers.” In *29th Pacific Asia Conference on Language, Information and Computation, PACLIC 2015*, , 267–74.
- Srivastava, Ashok, and Mehran Sahami, eds. 2009. *Text Mining: Classification, Clustering, and Applications*. CRC Press.
- Sundermann, Camila Vaccari, Marcos Aurélio Domingues, Merley da Silva Conrado, and Solange Oliveira Rezende. 2016. “Privileged Contextual Information for Context-Aware Recommender Systems.” *Expert Systems with Applications* 57: 139–58. <http://www.sciencedirect.com/science/article/pii/S0957417416301312>.
- Toshevskaa, Martina, Frosina Stojanovska, and Jovan Kalajdjieski. 2020. “Comparative Analysis of Word Embeddings for Capturing Word Similarities.” In *6th International Conference on Natural Language Processing (NATP 2020)*,

, 259–78.

Vaswani, Ashish et al. 2017. “Attention Is All You Need.” *Advances in Neural Information Processing Systems* 2017-Decem(Nips): 5999–6009.

Wang, Donghui et al. 2018. “A Content-Based Recommender System for Computer Science Publications.” *Knowledge-Based Systems* 157: 1–9. <http://www.sciencedirect.com/science/article/pii/S0950705118302107>.

Wang, Peng et al. 2015. “Semantic Clustering and Convolutional Neural Network for Short Text Categorization.” In *53rd Annual Meeting of the Association for Computational Linguistics and 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, , 352–57.

———. 2016. “Semantic Expansion Using Word Embedding Clustering and Convolutional Neural Network for Improving Short Text Classification.” *Neurocomputing* 174: 806–14.

Xie, Junyuan, Ross Girshick, and Ali Farhadi. 2016. “Unsupervised Deep Embedding for Clustering Analysis.” *33rd International Conference on Machine Learning, ICML 2016* 1: 740–49.

xu, jiaming et al. 2015. “Short Text Clustering via Convolutional Neural Networks.” : 62–69.

Xu, Yuanbo et al. 2019. “Slandorous User Detection with Modified Recurrent Neural Networks in Recommender System.” *Information Sciences* 505: 265–81. <http://www.sciencedirect.com/science/article/pii/S0020025519307017>.

Yang, Bo, Xiao Fu, Nicholas D Sidiropoulos, and Mingyi Hong. 2017. “Towards K-Means-Friendly Spaces: Simultaneous Deep Learning and Clustering.” In *International Conference on Machine Learning*, PMLR, 3861–70.

Yang, Jianwei, Devi Parikh, and Dhruv Batra. 2016. “Joint Unsupervised Learning of Deep Representations and Image Clusters.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, , 5147–56.

Yang, Shuiqiao, Guangyan Huang, and Borui Cai. 2019. “Discovering Topic Representative Terms for Short Text Clustering.” *IEEE Access* 7: 92037–47.

Zarinbal, M., M.H. H Fazel Zarandi, and I.B. IB B Turksen. 2014. “Relative Entropy Fuzzy C-Means Clustering.” *Information Sciences* 260(1): 74–97. <http://linkinghub.elsevier.com/retrieve/pii/S0020025513007895> (December 27, 2013).

Zarinbal, M, M H Fazel Zarandi, and I B Turksen. 2015. “Relative Entropy Collaborative Fuzzy Clustering Method.” *Pattern Recognition* 48(3): 933–940. <http://dx.doi.org/10.1016/j.patcog.2014.09.018>.

Zhai, Cheng Xiang, and Sean Massung. 2016. *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. ACM.

Recommender systems help users to make decisions based on their informational needs. These needs are recognized by users' queries, users' selected documents, and users' behavior. Knowledge based recommender systems or content based recommender systems are two well-used systems, in which machine learning and pattern recognition is applied. Text clustering is an important approach in machine learning that uses documents features and similarities to other documents to cluster them into meaningful classes. Thus, determining features and similarity measures are two most important task in any clustering method. On the other hand, almost all of the clustering methods are developed for non-Farsi languages. Therefore, the aim of this research is to develop a clustering method for scientific documents in Farsi. A set of documents of Ganj database in three different subject areas (Art, Humanities and Engineering) is used and three similarity measure including Minkovisky, Cosine and Entropy, is selected. K-means, FCM, and REFCM clustering algorithms are applied on the dataset. The results show that FCM with Cosine similarity and REFCM could achieve to better results.

Keywords: Scientific documents; Machine learning; Deep clustering; Relative Entropy



**Iranian Research Institute
for Information Science and Technology
(I r a n D o c)**

Research Project

**Developing a clustering approach to
recommend Persian scientific
documents based on Persian data in
Iran Scientific Information Database
(Ganj)**

By:

Marzieh Zarinbal Masouleh

Azadeh Mohebi

February 2022