Information Extraction

استاد محترم: جناب آقای دکتر غایبی

گردآورنده: آمنه شناور

دوره دکتری ورودی ۱۳۹۹

گروه دکتری بازیابی اطلاعات و دانش

مقدمه

- امروزه وب جهان گستر به علت توزیع شدگی و هزینه پایین تولید محتوا با چالش های جدیدی از جمله حجم زیاد اطلاعات، ناهمگنی وغیرساختاریافته بودن اطلاعات مواجه شده است.

- اطلاعات غیرساخت یافته قابل خواندن، سازماندهی و تحلیل توسط ماشینها نیستند.

- برای اینکه بتوان از بین این حجم عظیم اطلاعات، انسان را در فهم و یافتن اطلاعات مورد نیاز یاری کرد باید بتوان متن غیرساخت یافته را به اطلاعات ساخت یافته تبدیل کرد.

- در واقع نیاز به سیستمی وجود دارد که بتواند داده ها را به شکل ساخت یافته درآورد. استخراج اطلاعات شامل توسعه الگوریتم هایی است که بصورت خودکار، متن غیرساخت یافته را پردازش و پایگاه داده ای از موجودیتها، روابط و وقایع را تولید میکنند.

استخراج اطلاعات عبارتست از استخراج حقایق از درون اسناد غیر ساخت یافته و ساخت یافته. این حقایق اشیاء ساخت یافته مانند رکوردی از پایگاه داده هستند که این رکورد می تواند یک موجودیت در دنیای واقعی و ویژگی های مربوط به آن در متن مذکور باشد یا یک اتفاق در دنیای واقعی یا یک وضعیت به همراه پارامترها و بازیگران آن باشد

(مثلا چه کسی، چه کاری را برای چه کسی انجام داد، کی و کجا)

مقدمه

- استخراج اطلاعات یکی از برجسته ترین تکنیک هایی ست که در متن کاوی مورد استفاده قرار می گیرد.
- به عنوان اولین گام، در تَگ کردن و برچسب دهی به اسناد و مدارک در نظامهای تحلیل متن، هر سند و مدرک به منظور مشخص کردن موجودیتها و روابط معنادار بین آنها، باید پردازش شود.
- اصطلاح روابط در اینجا به حقایق یا رویدادهای مرتبط اشاره می کند.

به عنوان مثال، یک رویداد احتمالی ممکن است ورود یک شرکت به یک ریسک مشترک برای توسعه یک داروی جدید باشد.
یک مثال از یک واقعیت این است که یک ژن باعث بیماری خاصی می شود. حقایق ثابت هستند و معمولاً تغییر نمی کنند.
رویدادها پویاتر هستند و به طور کلی دارای یک مهر زمانی خاص هستند.

- روشهای استخراج اطلاعات، امکان استخراج اطلاعات واقعی در متن و نه مجموعه محدودی از برچسبهای مرتبط با اسناد را فراهم می کنند.
- بنابراین تکنیک های پیش پردازش، که شامل استخراج اطلاعات میباشند، قصدشان بر این است که مدلهای غنی تر و انعطاف پذیرتری را برای اسناد و مدارک در نظامهای تحلیل متن ایجاد کنند.

- استخراج اطلاعات را می توان به عنوان شکل محدودی از "درک کامل متن" دید. هیچ تلاشی برای درک کامل سند در دست انجام نشده است.
- درعوض، انواع اطلاعات معنایی که باید از سند استخراج شوند، تعریف می شود.

- **IE** اسناد را به عنوان مجموعه ای از موجودیت ها و چارچوب ها نشان می دهد که روش دیگری برای توصیف رسمی روابط بین موجودیت ها هستند.

مجموعه تمام موجودیت ها و چارچوب های ممکن در مقایسه با مجموعه کلیدواژه های دسته بندی معمولاً باز و بسیار بزرگ است.نمی توان آن را به صورت دستی ایجاد کرد.در عوض، ویژگی ها مستقیماً از متن استخراج می شوند.

ساده ترین نوع استخراج اطلاعات، استخراج اصطلاح است.
تنها یک نوع موجودیت وجود دارد . *“term.”*

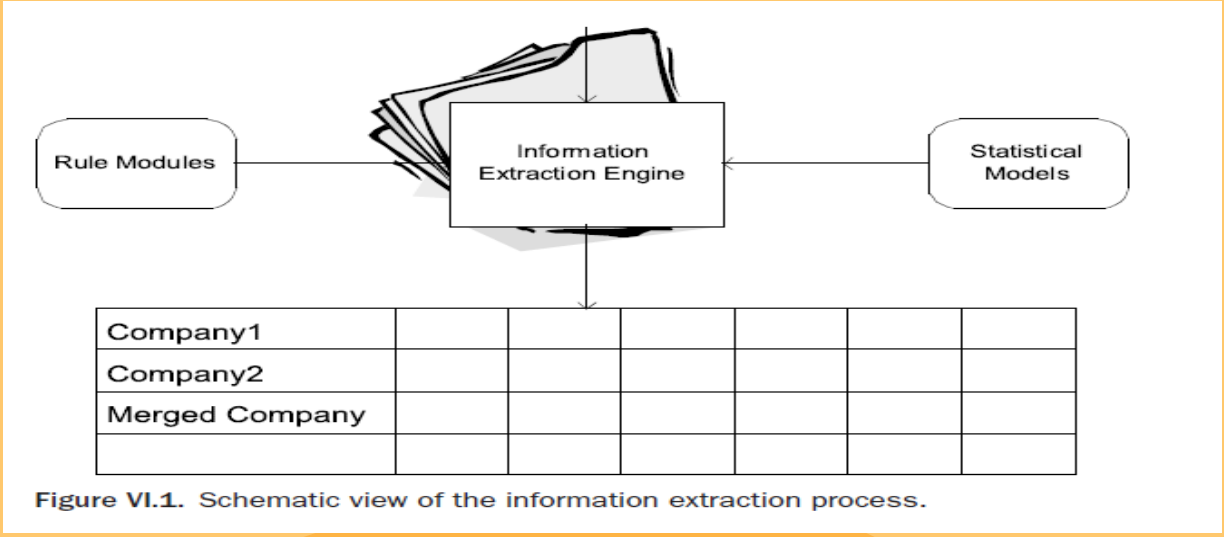
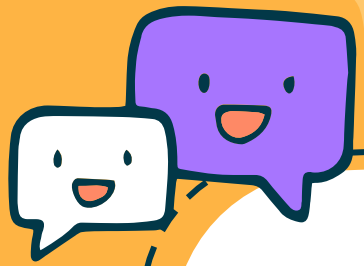


Figure VI.1. Schematic view of the information extraction process.

نظام استخراج اطلاعات

در مرکز فرآیند ما موتور استخراج اطلاعات را داریم که مجموعه ای از اسناد را به عنوان ورودی می گیرد.
موتور با استفاده از یک مدل آماری، یک ماژول قانون یاترکیبی از هردو کار می کند.
خروجی موتور مجموعه ای از چارچوبهای حاشیه ای که از اسناد استخراج شده است.
چارچوب ها در واقع جدولی را پر می کنند که در آن فیلدهای چارچوب ردیف های جدول هستند.

استخراج اطلاعات



"اشاره به استخراج خودکار اطلاعات ساختار یافته همچون موجودیت ها، روابط بین موجودیتها، و موجودیتهای توصیف ویژگی از منابع غیر ساختاریافته دارد."

بنابراین استخراج اطلاعات،
شناسائی مفاهیم از پیش تعریف شده و نادیده گرفتن اطلاعات بی ربط است

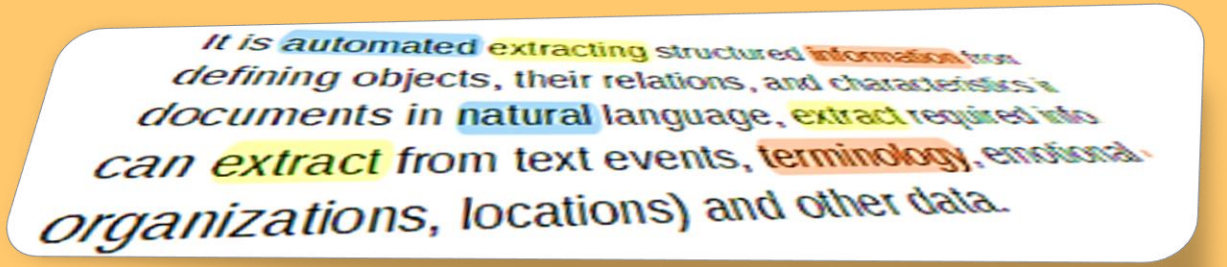
استخراج اطلاعات – ادامه

راه های جدیدی را برای جستجو، سازماندهی و تحلیل داده ها با بهره گیری از علم معناشناسی از پایگاه داده های ساختاریافته و داده های غیرساختاریافته گشوده است.

بخش زیادی از فعالیتهای استخراج اطلاعات مربوط به پردازش متون توسط روش پردازش زبانهای طبیعی است.

به طور خاص ، در این فن آوری میتوان با استفاده از ترکیب روش ها و تکنیک های پردازش زبان طبیعی کارایی بالایی را برای کاوش متن از دامنه های گوناگون به دست آورد .

بنابراین، استخراج اطلاعات ، یک فن آوری منشعب از پردازش زبان طبیعی ست، که به تحلیل متن فاقد ساختار و به زبان طبیعی برای مشخص نمودن اطلاعات یا وقایعی می پردازد که به صورت صریح یا ضمنی در آن وجود دارد.



Historical Evolution

- در حالی که از دهه ۱۹۵۰ میلادی که برای اولین بار ایده استخراج اطلاعات توسط دانشمندی آمریکایی به نام زلاهریس برای تبدیل خلاصه تاییدیه بیماران به ساختار جدولی در بیمارستان مورد استفاده قرارگرفت.
- اما اهمیت موضوع با شروع کنفرانس های درک پیام(Muc-6,Muc-7) و شروع تامین اعتبار ملی DARPA با نام ACE از سال ۱۹۹۹ با هدف استخراج خودکار محتوا از متون به زبان طبیعی تا به امروز ابعاد مختلفی از استخراج اطلاعات مورد توجه دانشمندان قرار گرفته است.

MUCهای اولیه به صورت الگوهای از پیش تعریف شده شامل یک سری شیار از پیش تعریف شده بودند.

یکی از مشکلات مربوط به وظیفه پر کردن شکاف این است که به شدت به برنامه وابسته است و چنین سیستم هایی در تنظیمات مختلف برنامه قابل تعمیم نیستند.

برای مثال، یک سیستم پر کردن شکاف برای سیاستمداران ایالات متحده در ویکی‌پدیا ممکن است برای پر کردن شکاف‌های تروریسم محور در مقالات خبری کار نکند.

مشکل اصلی این است که تنظیمات مختلف یا به سفارشی سازی قابل توجهی نیاز دارند، یا به پیچیدگی قابل توجهی از نظر نحوه ورودی نیاز دارند.

Slot/Field	Value
Born	May 29, 1917
Political party	Democratic
Spouse(s)	Jacqueline Bouvier
Parents	Joseph Kennedy Sr. Rose Kennedy
Alma mater	Harvard University
Positions	US House of Representatives US Senate US President
Military Service	Yes

Historical Evolution

یکی از سری کنفرانس های درک پیام بود. مشخص شد که وظایف واضح‌تری مانند شناسایی موجودیت نام‌گذاری شده و استخراج رابطه اغلب به عنوان وظایف فرعی پر کردن شکاف استفاده می‌شوند،

و همچنین از مزیت مستقل بودن الگو برخوردارند.

در نتیجه، این وظایف فرعی در نهایت به اشکال غالب استخراج اطلاعات تبدیل شدند.

معنادار کردن کلمات، بندها یا ترکیبی از انها در متن می باشد.
لذا فرایند معنادار کردن می تواند به صورت تبدیل متن به یک فایل جدول یا به صورت یک متن حاشیه نویسی شده باشد.

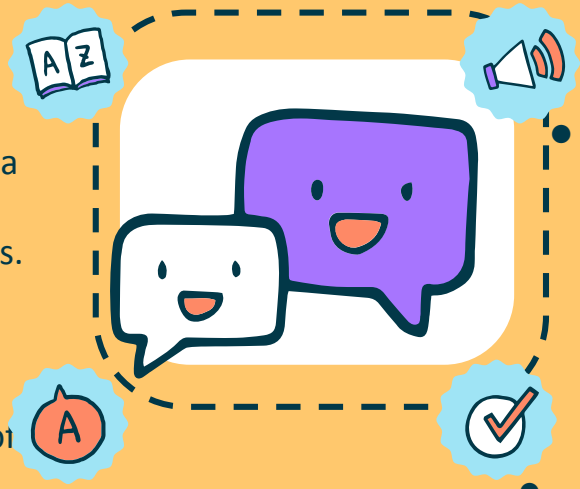
Elements That Can Be Extracted from Text

Facts

- Facts are the relations that exist between entities. Some examples are an employment relationship between a person and a company or phosphorylation between two proteins.

Events

- An event is an activity or occurrence of interest in which entities participate such as a terrorist act, a merger between two companies, a birthday and so on.



Entities

Entities are the basic building blocks that can be found in text documents. Examples include people, companies, locations, genes, and drugs

Attributes

Attributes are features of the extracted entities. Some examples of attributes are the title of a person, the age of a person, and the type of an organization.

we can extract the following information from the text: Example

Indian captain Virat Kohli was dismissed cheaply for just 2 in Wellington on Friday by debutant Kyle Jamieson extending a rare lull in the batsman's stellar career. Throughout the ongoing New Zealand tour, Kohli has managed to score just a single fifty across 8 innings in all 3 international formats.

- Country – India, Captain – Virat Kohli
- Batsman – Virat Kohli, Runs – 2
- Bowler – Kyle Jamieson
- Match venue – Wellington
- Match series – New Zealand
- Series highlight – single fifty, 8 innings, 3 formats

Marc Marquez was fastest in the final MotoGP warm-up session of the 2016 season at Valencia, heading Maverick Vinales by just over a tenth of a second.

After qualifying second on Saturday behind a rampant Jorge Lorenzo, Marquez took charge of the 20-minute session from the start, eventually setting a best time of 1m31.095s at half-distance.

Person: Marc Marquez

Location: Valencia

Event: MotoGP

Related mentions: Maverick Vinales, Yamaha, Jorge Lorenzo

مثالی از استخراج اطلاعات

19 March – a bomb went off this morning near a power tower in San Salvador leaving a large part of the city without energy, but no causalities have been reported. According to unofficial sources, the bomb – allegedly detonated by urban guerrilla commandos – blew up a power tower in the northwestern part of San Salvador at 0650 (1250 GMT).

Incident Type:	Bombing
Date:	March 19th
Location:	El Salvador: San Salvador (City)
Perpetrator:	urban guerrilla commandos
Physical Target:	power tower
Human Target:	–
Effect of Physical Target:	destroyed
Effect on Human Target:	no injury or death
Instrument	bomb

1 Named entity recognition

نشانه های موجود در متن ممکن است به موجودیت های نامگذاری شده اشاره داشته باشند، مانندمکان ها، افراد و سازمان ها

Bill Clinton lives in New York at a location that is a few miles away from an IBM building. Bill Clinton and his wife, Hillary Clinton, relocated to New York after his presidency.

"نیویورک" یک مکان است،
"بیل کلینتون" یک شخص است
و " "IBMیک سازمان است.

2 Relationship extraction

تشخیص و طبقه بندی روابط از پیش تعریف شده بین موجودیت های مشخص شده در متن

- رابطه بین یک فرد و یک سازمان

Steve Jobs works for Apple
Employee of (Steve Jobs ,Apple)

- رابطه بین یک فرد و محل

Mr. Smith gave a talk at the conference in New York
Located In (Smith ,New York)

- رابطه بین دو شرکت

Listed broadcaster TVN said its parent company, ITI Holdings, is considering various options for the potential sale.
Subsidiary Of (TVN ,ITI Holding)

ARCHITECTURE OF IE SYSTEMS

توکنایزیشن یا تقسیم بندی و قطعه بندی

مشخص واحدهای معنایی است که یک سند یا مدرک ورودی را به سازه های اصلی آن تقسیم بندی میکند. سازه های اصلی معمولاً کلمات، جملات و بندها (پاراگرافها) هستند و به ندرت ممکن است بلوک ها یا سازه هایی مانند بخش ها یا فصل ها را داشته باشیم

پردازش واژگانی و ریخت شناسی

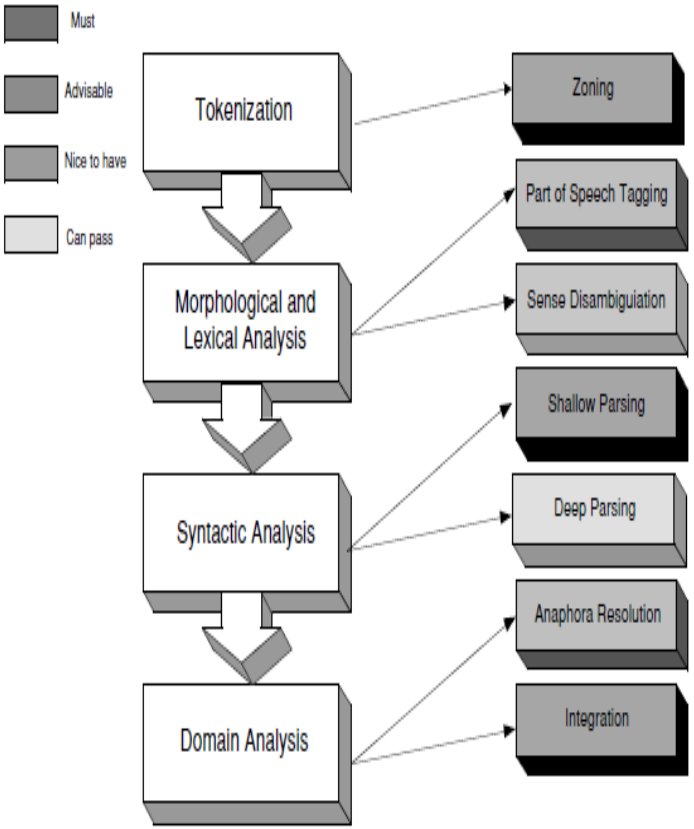
بر فعالیتهایی مانند اختصاص برچسب های پارت آف اسپیچ تگینگ یا (POS)(تحلیل اجزای کلام و جمله و گفتار و مشخص کردن نقش گرامری کلمات) به کلمات گوناگون اسناد و مدارک، ایجاد عبارت های اساسی مانند عبارت های اسمی و فعلی و رفع ابهامَ از کلمات و عبارت ها تمرکز میکند.

تحلیل نحوی

ارتباط بین بخشهای گوناگون هر جمله را فراهم میکند. این فرایند به وسیله تحلیل کامل و عمیق فراهم میشود

تحلیل دامنه

نظام، تمامی اطلاعات جمع آوری شده از اجزای پیشین را ترکیب میکند و یک چهارچوب کامل برای توصیف ارتباطات بین موجودیت ها فراهم میکند.



معماری نظام استخراج اطلاعات

شناسایی موجودیتهای اسمی، یک زیروظیفه از استخراج اطلاعات است.

شناسایی موجودیتهای نامدار (NER) در پردازش زبان طبیعی به عملیاتی گفته میشود که در طی آن کلیه ی اسامی خاص موجود در متن و متعلق به مقوله های معنایی مختلف،شناسایی و استخراج میگردند و تحت کلاس‌های از پیش تعریف شده ای مانند اسم افراد، سازمان‌ها، مکان‌ها و ... دسته‌بندی می‌شوند. به این صورت که متن را براساس واژگان قطعه بندی و عبارات حاوی موجوودیت نامدار را با برچسب زنی مشخص می کنیم.

شناسایی موجودیت نام‌گذاری شده اساسی‌ترین مشکل در استخراج اطلاعات است،
زیرا بلوک اصلی را فراهم می‌کند که بسیاری از روش‌های استخراج اطلاعات دیگر بر روی آن ساخته شده‌اند.
برای مثال، انجام استخراج رابطه غیرممکن خواهد بود، اگر موجودیت‌های نام‌گذاری شده برای استخراج روابط وجود نداشته باشد.

برای تشخیص موجودیت های نامدار دو روش

- ✓ Rule-Based Methods
- ✓ statistical learning methods

روشهای قاعده پایه

هر نشانه در متن به مجموعه ای از ویژگی ها تبدیل می شود. این ویژگی‌ها معمولاً ویژگی‌های مفید (Token) یا زمینه آنها برای استخراج موجودیت هستند.

به عنوان مثال، یکی از ویژگی های واضح می تواند اطلاعاتی در مورد اینکه آیا آن نشانه با حرف بزرگ شروع می شود یا نه. بنابراین، این ویژگی ها به تعریف الگوهای مختلف در قانون کمک می کند.

قالب و شکل کلی موجودیتهای اسمی را به صورت یک عبارت باقاعده نمایش داده شده تا سامانه اسامی خاص را بر مبنای این عبارات تشخیص دهد.

دو قانون اصلی الگوریتم های یادگیری این سیستم ها

روش پایین به بالا

که موارد را از خاص به عمومی فرا میگیرد.

LP2

یکی از انواع روش ها ی پا یین به بالا است. که دو نوع قانون را آموزش میدهد.که به ترتیب مرز آغازین و مرز پا یانی متنی که استخراج شده است را شناسایی میکند،آموزش از مثال ها در یک مجموعه تعریف شده توسط کاربر انجام شده است.(مجموعه داده های اولیه (آموزشی)).آموزش در دو مرحله انجام شده است:

۱- در ابتدایک مجموعه از قوانین برچسب زنی آموزش داده می شود.

۲-قوانین اضافی جهت تصحیح اشتباهات و بی دقتی در استخراج به وجود آمده اند.

روش بالا به پایین

که موارد را از عمومی به سمت خاص فرا میگیرد.

WHISK

الگوریتم آن به این صورت است که از عمومی ترین قانون شروع کرده و به صورت تصاعدی با اختصاصی کردن قوانین موجود ادامه میدهد. الگو ریتم هنگامی خاتمه می یابد که قانونی ایجاد شده باشد که تمام مثال های مثبت آموزشی را پوشش داده باشد.

استخراج ویژگی

ویژگی‌های کلمه‌ای

این ویژگی ها مربوط به نویسه های سازنده ی کلمه است و از روی شکل وظاهر خود کلمه استخراج می گردد.
ویژگی هایی که مربوط به بزرگ یا کوچک بودن حروف در کلمه

ویژگی‌های فهرستی

برای استخراج این ویژگی ها از فهرست هایی که شامل اسامی افراد ، مکان ها ، شهرها ، کشورها ، افراد و ... است استفاده می شود

ویژگی های سندی

این ویژگی ها مربوط به اطلاعاتی از کلمه است که در کل سند وجو دارد . این اطلاعات از طریق پردازش کل اسناد به دست می آید . واضح است که اگر حجم اسناد و داد های ما زیاد باشد ، ویژگی های قوی تری استخراج می‌گردد.

روشهـــای آمـــاری و یادگیـــری ماشـــین

در سیستم های مبتنی بر یادگیری ماشینی ،هدف از رهیافت تشخیص واحدهای اسمی تبدیل مساله تشخیص به مساله دسته بندی است و از یک مدل آماری دسته بندی برای حل این مساله استفاده می شود

.در این روش مدل به دنبال تشخیص الگوها و یافتن رابطه ی آنها با متن و ساختن یک مدل آماری و الگوریتم یادگیری ماشین است. این سیستم ها نام ها را یافته و آن ها را بر اساس مدل به دست آمده با استفاده از روش های یادگیری ماشین به کلاس های از پیش تعیین شده مانند اشخاص ، مکان ها ، زمان ها و ... تقسیم می کند. این مدل یادگیری، یادگیری با ناظر است یعنی سیستم با استفاده از یک مجموعه از مثال های برچسب گذاری شده ،دسته بندی را یاد می گیرد. در این روش ابتدا سامانه به وسیله ی پیکره ای از داده های آموزشی که به صورت دستی و به وسیله ی انسان برچسب گذاری شده اند آموزش دیده، با یادگیری از طریق این داده ها به تشخیص خودکار اسامی خاص در متن می پردازد، که در بخش ویژگی های سندی اشاره شد.

روشهای مبتنی بر یادگیری که به صورت با نظارت عمل می کنند،نیاز به حجم زیادی از متون برچسب گذاری شده دارند.

برخلاف رهیافتهای مبتنی بر قاعده، روشهای یادگیری ماشین مستقل ازحوزه و زبان عمل می کنند.

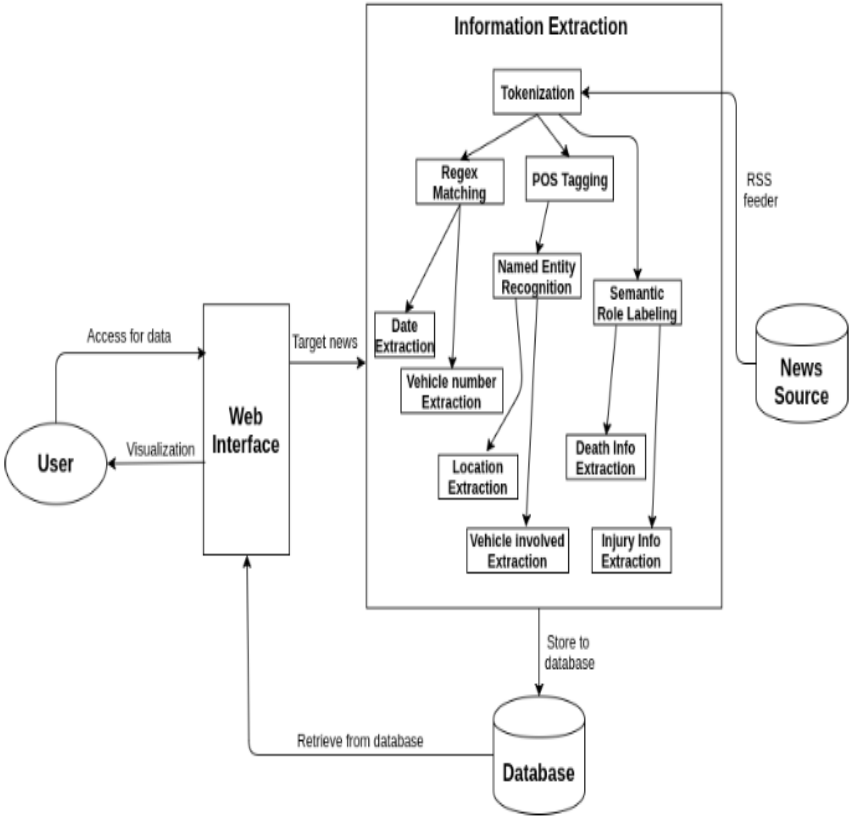
زمینه های تصادفی شرطی

مدل های مارکوف حداکثر آنتروپی

مدل های پنهان مارکوف

Applications information extraction

- خلاصه سازی
- داده‌های زیست‌پزشکی
- سیستم های پرسش و پاسخ
- بازیابی اطلاعات
- تقسیم بندی متن



Information Extraction from News Article and Its Analysis

انواع استخراج اطلاعات

هدفمند

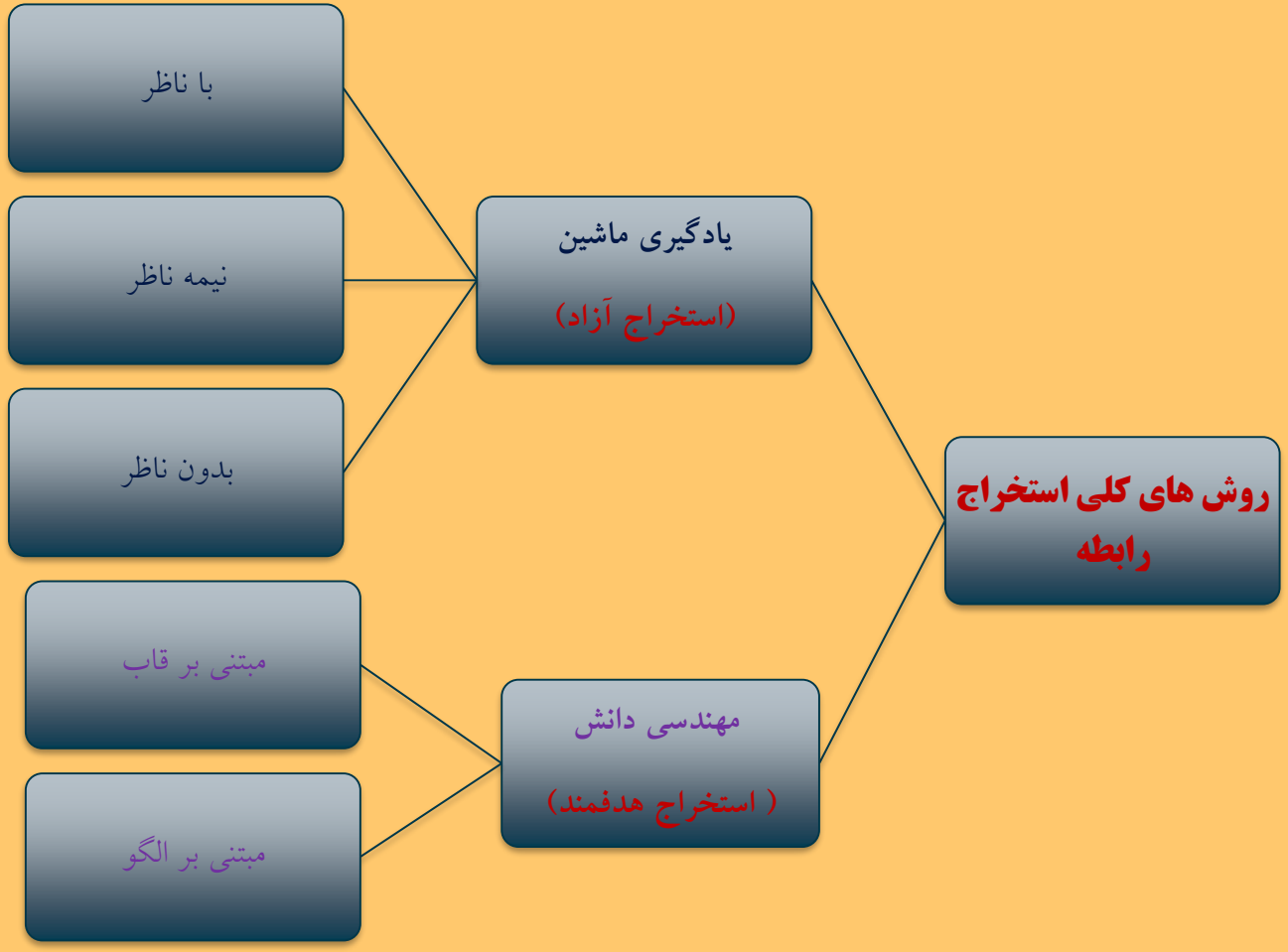
زمانی که اطلاعات مشخصی برای استخراج، توسط انسان معین شده اند و حالا ماشین باید این داده ها را یاد بگیرد تا بتواند از متون جدید نیز اطلاعات مورد نظر را استخراج کند در واقع استخراج هدفمند را برگزیده ایم.

مثلا استخراج زمان و مکان برگزاری کنفرانس ها

آزاد

در استخراج آزاد اطلاعات با تعدادی نمونه آغازین، مربوط به هر رابطه یا بدون هیچ داده ای برای آموزش، اطلاعات را استخراج میکنند. استخراج آزاد اطلاعات، یکی از روشهای استخراج رابطه است.
برخلاف روشهای پیشین استخراج اطلاعات، استخراج همه روابط دلخواه ازجملات موجود در متن را فراهم میکند.

استخراج نمونه های رابطه در متون بزرگ مانند وب


روش های کلی استخراج رابطه
یادگیری ماشین
(استخراج آزاد)
با ناظر
نیمه ناظر
بدون ناظر
مهندسی دانش
(استخراج هدفمند)
مبتنی بر قاب
مبتنی بر الگو

روشهای یادگیری ماشینی

سامانه های با ناظر (خودناظر)

سامانه باناظر، داده آموزشی را برای خود تهیه میکند. تهیه نمونه برای هر رابطه، اگر چه تعداد این داده ها بسیار کم باشد، (روشهای باناظر با داده آموزشی کم کار میکنند)،

سامانه را با چالش مناسب بودن داده ها مواجه میکند. یعنی باید مطمئن باشیم که داده فراهم شده، موجب بازیابی بخش خاصی از نتایج نمیشود. در مقابل این رویه، سامانه های باناظر با ادعای حذف نمونه های آغازین تلاش میکنند دادههای موردنیاز خود را تولید کنند و میتوان آنها را جزو سامانه های بی ناظر به شمار آورد. این روش به جای داده های آغازین با تعدادی الگو و نوع شروع میشود. قرار دادن هر نوع در الگوها، به ایجاد جستجوی مناسب برای بازیابی نمونه های آن نوع میشود. یعنی نمونه ها از طریق انطباق الگو با نتایج جستجو استخراج میشوند.

- دسته بندی براساس ویژگی
- روش هسته

روشهای یادگیری ماشینی

سامانه های با ناظر (خودناظر)

- **دسته بندی براساس ویژگی**

به طور خاص، هر جفت موجودیتی که درجمله اتفاق میافتد بعنوان نامزد مطرح میشود. هدف تخصیص برچسب کلاس به جفت موجودیت است که برچسب کلاس یک رابطه از پیش تعریف شده برای جفت موجودیت نامرتبط است.
مهندسی ویژگی گام مهم در روش دسته بندی است. ویژگیهای موجودیت ، ویژگیهای متنی واژگانی، ویژگیهای متنی نحوی

- **روش هسته**

در یادگیری ماشینی، یک تابع هسته یا کرنل محصول داخلی نمونه های مشاهده شده را در بعضی زیر لایه های فضای برداری تعریف میکند . مزیت عمده ی استفاده از هسته این است که موارد مشاهده شده برای محاسبه شدن لازم نیست به صراحت به فضای برداری محصولات داخلی خود نگاشت شود.

روشهای یادگیری ماشینی

سامانه های بی ناظر

سامانه های بی ناظر به داده های آموزشی نیازی ندارند و سعی در تولید آنها ندارند. نام روش بی ناظر، رابطه محکمی با خوشه بندی دارد و میتوان آن را رکن اصلی سامانه هایی دانست که به صورت بی ناظر استخراج میکنند.

در واقع ما دنبال یافتن ارتباط بین داده ها و برچسب مربوط به آن نیستیم. در این حالت فقط سعی می شود که داده هایی که شبیه به هم هستند داخل یک گروه قرار بگیرند.

این سامانه ها با خوشه بندی اخبار براساس مدل رخدادها (Bag of Words)، خبرهای مربوط به یک اتفاق با زمان مشخص را در یک دسته قرار میدهند. سپس با خوشه بندی دسته های حاصل، براساس الگوهای نحوی موجودیتهای اسامی آنها، خوشه های وقایع را ایجاد میکند.

در این روش، ماشین، داده ها را بر اساس شباهت ها، الگوها و تفاوتها گروه بندی میکند و سعی در یافتن الگوهای پنهان موجود در داده ها دارد.

حادثه ای مثل وقوع طوفان با الگوهای سرعت طوفان و میزان تلفات شهر شناخته میشود.

روشهای یادگیری ماشینی

سامانه های‌نیمه ناظر

سامانه های نیمه ناظر مشکل تهیه نمونه های آغازین را ندارند. ظهورسامانه هایی که با استفاده از چند نمونه آغازین، الگوهای وقوع اطلاعات را یاد میگیرند و آنها را استخراج میکنند، باعث شد که زحمت انسانی لازم برای تعریف قوانین استخراج کم شود.

✓ روش خودراه انداز

روش مهم در یادگیری نیمه ناظر روش خود راه انداز است که از مجموعه کوچک نمونه های رابطه شروع میشود و از الگوهای استخراج استفاده میکند.

سامانه اسنوبال :ایده این سامانه ساده است و با جفت موجودیتهای مرتبط با رابطه هدف شروع میکند و در متن به جستجوی جفت موجودیتهای مجاور هست،اگر جفت موجودیتها بطور همزمان در متن رخ داده باشند، مفهوم همزمانی موجودیتها احتمالا به معنی الگویی برای رابطه هدف است. پس جفت موجودیتها به نمونه رابطه اضافه میشوند و تا زمانیکه شرایط دقیقی ایجاد شود پردازش ادامه دارد بطوریکه بیشتر الگوها و موجودیتها به نتایج پردازش اضافه میشوند . یک گام مهم در شیوه خودراه انداز ارزیابی کیفیت الگوهای استخراج است درنتیجه فرآیند استخراج شامل الگوهای خراب نمیشود.

✓ روش نظارت راه دور

روش نظارت راه دور ویژگیهای استخراج شده از جملات متفاوت شامل هر جفت موجودیت را برای ایجاد بردار ویژگی غنی استفاده میکند.

روش نظارت راه دور یا یادگیری خودنظارتی برای استخراج پایگاه های‌دانش بزرگ برای برچسب زدن خودکار موجودیتهای در متن و استخراج ویژگیهاو آموزش دادن دسته بند بکار میرود. این روش فقط برای استخراج روابطی که از مرز جملات رد نشده اند و جملاتی که حاوی اشاره روشنی از فعل و فاعل رابطه است استفاده میشوند. استخراج نظارت راه دور بعنوان برچسب گذاری خودکار متن با خصوصیات و منابعی که منابع موجودیتها ا ز یک پایگاه دانش هستند استفاده میشوند.

کاربردهای استخراج آزاد اطلاعات

استخراج دانش، استنتاج از زبان طبیعی، استخراج خودکار از شبکه معنایی، پاسخ به پرسش ها و مدل زبانی رابطه ای

استخراج دانش

به عنوان اولین مرحله استخراج دانش از متن، نیاز است تا موجودیتهای متن را استخراج کرده و نام های هم معنا که به یک موجودیت مربوط میشوند، با هم در یک گروه قرار دهیم. البته هر نام ممکن است در چند گروه قرار گیرد. برای نمونه نام"حسن روحانی" میتواند در گروههایی متفاوت همراه با "رئیس جمهور" و "سیاست مدار: و"استاد دانشگاه" قرار گیرد. این کار بسیار مشابه با مسئله یادگیری "هستان شناسی "خواهد بود.

روشهایی وجود دارد که برای این کار از منابع دانش خارجی نظیر وردنت یا ویکی پدیا استفاده میکنند که به عنوان معروف ترین آنها میتوان از **Wikitaxonomy** و **Yago** نام برد. یاگو با استفاده از وردنت و همچنین دسته های موضوعی ویکی پدیا، روشی برای تولید پایگاه دانش به طور خودکار ارائه کرده است.

استنتاج از زبان طبیعی

مسئله تشخیص این است که آیا یک جستجو که به زبان طبیعی بیان شده است، میتواند به طور منطقی از متون موجود که آنها نیز به زبان طبیعی بیان شده اند استخراج شود.

سامانه های استخراج آزاد اطلاعات

Textrunner

این سامانه با اعمال تعدادی قانون روی دادهها ،برای خود تعدادی نمونه صحیح ایجاد کرده و سپس آنها را یاد میگیرد. این روش را با نام روش خودناظر نامگذاری کرده اند. سپس از این ابزار برای استخراج رابطه از داده ها استفاده میشود.

Reverb

فعل های موجود در متن را مییابد و سپس رابطه متناسب با هر فعل را استخراج میکند. تجزیه کننده نحوی را برای برچسبگذاری جملات استفاده میکند و محدودیتهای واژگانی و نحوی را برای شناسایی واقعیات دودوئی بکار میبرد.

OLLIE

ابتدا مجموعه سطرهایی از سامانه ریورب را با خودراه انداز مجموعه آموزشی بزرگ به کار میبرد و قالبهای الگوی باز را روی این مجموعه آموزشی یاد میدهد که این قالبهای الگو در زمان استخراج به کار میروند.

WOE

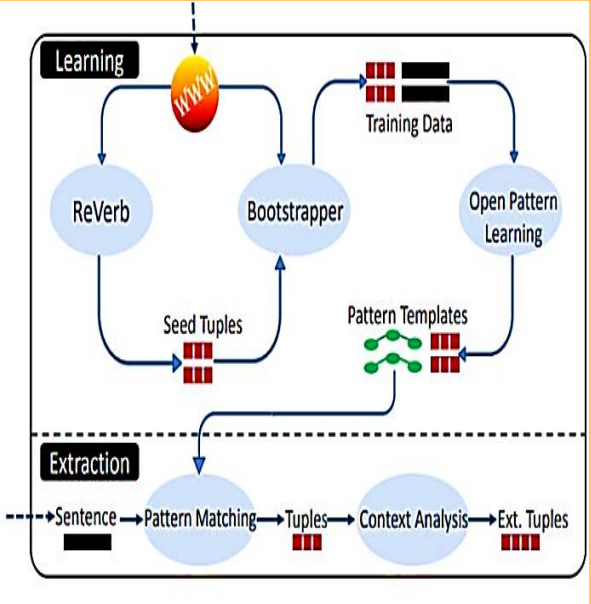
با استفاده از دادههای ساخت یافتها ی که در صفحات ویکی پدیا وجود دارد داده های مورد نیاز آموزش را ایجاد میکند این سامانه محدودیتهای معنایی–لغوی برای الگوها قرار نمیدهد و برای عبارتهای رابطه که فعل میانجی شده دارد و شامل اسم نیست طراحی شده است.

Kraken

برای گرفتن حقایق کامل از جملات عرضه شد و میتواند حقایق یک تایی، دوتایی تا چندتایی را استخراج کند.

Know-it-all

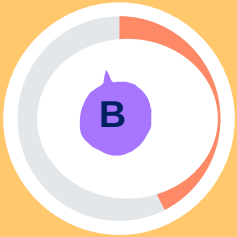
نوایت آل برای شروع کار خود به تعدادی الگو و شرح داده موردنظر برای استخراج نیاز دارد. این الگوها وابسته به زبان و البته مستقل از رابطه هستند. سامانه با استفاده از الگوها و داده موردنظر تعدادی عبارت تولید میکند و با استفاده از موتور جستجو، صفحات وب مربوط به آن را بازیابی میکند و درنهایت اطلاعات از این صفحات بازیابی شده استخراج میشود.



OLLIE's system architecture

چالشهای استخراج آزاد اطلاعات

سیستمها نیز قادر به استخراج تمام روابط نیستند و از طرفی خروجی ناقص و نوفه دار دارند و نیز ممکن است استخراج اطلاعی را در برنداشته باشند



بدلیل ماهیت مقیاس پذیر بودن استخراج آزاد روابط، استفاده از ابزارهای عمیق پردازش زبان طبیعی نظیر تجزیه گر نحوی و معنایی که باعث بهبود قابل توجه نتایج و افزایش دقت میشود، ممکن نیست.

استفاده صرف از ابزارهای سطحی پردازش زبان طبیعی نظیر تجزیه گر سطحی، اجزای سخن و .... باعث کاهش چشمگیری در معیارهای کارایی استخراج گرها میشود.

روشهای مهندسی دانش

روش مبتنی بر قالب

منظور از قالب، نحوه بیان یک واقعه به همان شکلی است که معمولاً بیان میشود. در بیان هر واقعه ای معمولاً تعدادی نقش معنایی در شکلهای متنوع و البته محدود ظاهر میشوند. برای نمونه در یک خبر مربوط به بمب گذاری، از نقش عامل بمب گذار، منطقه آسیب دیده و ... صحبت میشود. روشن است که هر قالب حجم بسیاری از اطلاعات را در خود جای میدهد و پیشنهاد اقتباس و استفاده از آن به شکل بی ناظر کمی عجیب به نظر میرسد. استخراج اطلاعات به این نحو نیز تجربه شده است. ویژگیهای به دست آمده از متون براساس میزان باهم آیی آنها خوشه بندی میشود تا به خوشه هایی که هر کدام در مورد موضوع مشخصی صحبت میکنند، دست یابیم. پس از این مرحله امیدواریم که مثلاً یک خوشه مربوط به اخبار بمب گذاری باشد و خوشه دیگر مربوط به اخبار آدم ربایی ، متون هم موضوع در یک خوشه قرار میگیرند تا برای هر کدام از آنها قالبی با نقش های معنایی مشخص، کشف شود. یادآوری میشود که فرایند کاملاً بی ناظر و مبتنی بر دانش است.

روش مبتنی بر الگو

در روشهای مبتنی بر الگو، ورودی (معمولاً متن)، به دنبال الگو یا کلمه کلیدی ویژه که نشانگر رابطه مفهومی خاص است، جستجو میشود. این الگوها انواع مختلفی (اعم از نحوی یا معنایی و عمومی یا خاص) دارند و برای استخراج عناصر مختلف هستانشناسی مثل روابط طبقه ای یا غیرطبقه ای یا اصول بدیهی به کار میروند.

روشهای استخراج هدفمند اطلاعات

- به روشهای مبتنی بر قاعده که در سامانه هایی از قبیل [YAGO،] [DBpedia] از آن استفاده شده است. در این سامانه ها با استفاده از قواعد دست ساز، انبوه اطلاعات ساخت یافته موجود در ویکی پدیا و یا وردنت استخراج میشوند.

- روش دیگر استخراج هدفمند اطلاعات، استفاده از مدلهای گرافی است. برای نمونه استخراج ویژگیهای مقاله از میان سربرگ و ارجاعها به شکل یک مسئله پیش بینی ساختار تعریف و حل شده است .

- روش دیگر استخراج اطلاعات استفاده از توابع کرنل است که برای این کار تعریف شده و مورد استفاده قرار گرفته اند. برای نمونه استفاده از تجزیه کم عمق جمله برای تشخیص رابطه اشخاص و نهادها و همچنین مکان سازمانها بررسی شده است.

کاربردهای استخراج هدفمند اطلاعات

کاربردهای تجاری

- ✓ پیگیری اخبار
- ✓ مراقبت از مشتری
- ✓ تمایز داده ها
- ✓ تبلیغات طبقه بندی شده

مدیریت اطلاعات شخصی

- ✓ مدارک
- ✓ پیام نگارها
- ✓ پروژه ها
- ✓ افراد

کاربردهای علمی

- ✓ انفورماتیک زیستی

کاربردهای وب محور

- ✓ وبگاه های جوامع
- ✓ پایگاه داده های نظریه ای
- ✓ خرید مقایسه ای
- ✓ پایگاه های استنادی

نوع استخراج اطلاعات	روش استفاده شده	نوع دسته‌بندی		ویژگی‌های اصلی	رویکردها
یادگیری ماشین	باناظر	دسته‌بندی براساس ویژگی	موجودیت متنی واژگانی متنی و نحوی	✓ استفاده از داده آموزشی زیاد	استخراج هدفمند اطلاعات
		دسته‌بندی براساس هسته	هسته مبتنی بر ترتیب هسته مبتنی بر درخت هسته ترکیبی		
	نیمه‌ناظر	خودراه‌انداز		✓ استفاده از نمونه رابطه و داده آموزشی کم	
		نظارت راه دور			
	بدون‌ناظر	کشف رابطه و القای الگو		✓ استفاده از خوشه‌بندی	
مهندسی دانش	مبتنی بر قالب	خوشه‌بندی متون با موضوع یکسان و استخراج قالب مشخص		✓ خوشه‌بندی و استخراج قالب	استخراج آزاد اطلاعات
	مبتنی بر الگو	استخراج الگو یا کلمه کلیدی خاص از متن استخراج عناصر مختلف هستان شناسی		✓ الگوهای نحوی و معنایی	

مقایسه روش‌های استخراج اطلاعات

منابع

ایمانی، محسن (۱۳۹۲). «ابهام زدایی و ارزیابی اطلاعات استخراج شده از متن زبان طبیعی».پایان نامه کارشناسی ارشد، دانشگاه علم و صنعت ایران، تهران.

حاصلی، داوود؛ ملوک السادات حسینی بهشتی و سمیه پاک نهاد (۱۳۹۵). استخراج اطلاعات: روشها و کاربردها. اولین کنفرانس بین المللی بازیابی تعاملی اطلاعات. قابل دسترس در: https://www.civilica.com/

حیدری، سمیه؛ وحیده رشادت(۱۳۹۷).«ارائه روشی جهت بهبود دقت خروجی سامانه های استخراج آزاد اطلاعات با استفاده از روشهای یادگیری ماشین».پایان نامه کارشناسی ارشد، موسسه آموزش عالی پویش.

حیدری، سمیه؛ زهره بنائیان؛ وحیده رشادت(۱۳۹٦). بررسی روشهای استخراج اطلاعات مبتنی بر یادگیری ماشین و مهندسی دانش. د.وین کنفرانس بین المللی پژوهش های دانش بنیان در مهندسی کامپیوتر و فناری اطلاعات.

خوشیان، ناهید؛ امیر غایبی (۱۳۹۷).«درآمدی بر استخراج اطلاعات و استخراج مفاهیم در داده کاوی و تفاوت این دو فرایند با یکدیگر (روشها، کاربردها و چالشها) با تأکید بر کاربرد داده کاوی در سازمانهای پلیسی و قضایی به منظور کشف جرایم». توسعه سازمانی پلیس، شماره ٦٦ ، صص ۱۱۱ –۱۵۸.

رشادت، وحیده؛ مریم حورعلی (۱۳۹۳).« مروری بر روشهای استخراج رابطه در یادگیری هستان نگار و استخراج اطلاعات ».همایش ملی مهندسی رایانه و مدیریت فناوری اطلاعات.

C. C. Aggarwal and C. Zhai, Eds., Mining Text Data. Boston, MA: Springer US, 2012.

منابع

Feldman, Ronen, Sanger, James. (2006). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.* Cambridge University Press, NewYork

Jiang J. (2012) Information Extraction from Text. In: Aggarwal C., Zhai C. (eds) Mining Text Data. Springer, Boston, MA. https://doi.org/10.1007/978-1-4614-3223-4_2.

J. Piskorski and R. Yangarber, “Information Extraction: Past, Present and Future,” 2013, pp. 23–49.

Sarawagi, S. (2008), "Information extraction", Foundations and trends in database, 1(3), 261-377.

M. Banko, “Open Information Extraction for the Web,” 2009.\

سوالات

سوال یک: عناصراستخراج شده از متن را نام ببرید؟ وبرای هریک مثالی بزنید.

موجودیتها : که سازه های اصلی هستند و میتوانند از اسناد و مدارک متن استخراج شوند، شامل افراد، شرکتها، داروها و ژن هاست.
ویژگیها : ویژگیها، موجودیتهای استخراج شده هستند. چند نمونه از ویژگیها عبارت اند از عنوان یک فرد، سن فرد و نوع سازمان.
حقایق : ارتباطاتی هستند که بین موجودیتها وجود دارند. برای نمونه، رابطه کاری بین شخص و شرکت یا فرایند فوسفوریلاسیون بین دو پروتئین.
رویداد : یک فعالیت یا رخداد است که در آن موجودیتها عبارت اند از: یک اقدام تروریستی، ادغام دو شرکت.

سوال دو: نظام استخراج اطلاعات دارای چند جزء اصلی می باشد؟

توکنایزیشن یا تقسیم بندی و قطعه بندی
پردازش واژگانی و ریخت شناسی
تحلیل نحوی
تحلیل دامنه

سوالات

سوال سوم: انواع استخراج اطلاعات را نام ببرید و هرکدوم را توضیح دهید.

استخراج هدفمند :زمانی که اطلاعات مشخصی برای استخراج، توسط انسان معین شده اند و حالا ماشین باید این داده ها را یاد بگیرد تا بتواند از متون جدید نیز اطلاعات مورد نظر را استخراج کند در واقع استخراج هدفمند را برگزیده ایم. زمانی که اطلاعات مشخصی برای استخراج، توسط انسان معین شده اند و حالا ماشین باید این داده ها را یاد بگیرد تا بتواند از متون جدید نیز اطلاعات مورد نظر را استخراج کند در واقع استخراج هدفمند را برگزیده ایم.

استخراج آزاد: در استخراج آزاد اطلاعات با تعدادی نمونه آغازین، مربوط به هر رابطه یا بدون هیچ داده ای برای آموزش، اطلاعات را استخراج میکنند. استخراج آزاد اطلاعات، یکی از روشهای استخراج رابطه است.
برخلاف روشهای پیشین استخراج اطلاعات، استخراج همه روابط دلخواه ازجملات موجود در متن را فراهم میکند.
به عبارت دیگر، در برخی موارد هدف کشف تمام حقایق مفید و برجسته موجود در متون بزرگ و متنوع از جمله وب است که به استخراج اطلاعات آزاد نیاز دارد