فهرست

صفحه ۱

با سلام خدمت دوستان ،

با مقاله ای درمورد توسعه ی کوری (جستجو) در خدم تتون هستم ....

صفحه ۲

درباره ی مقاله

در سال ۲۰۱۸ متشر شده ، در اسپرینگر

دارای پنج نویسنده می باشد و از دانشگاه های تونس و فرانسه در نوشتن اون همکاری کردن

صفحه ۳

چکیده (خلاصه)

گسترش کوری چیست ؟

پروسه ی مهمی در کاربرد های بازیابی داده میباشد

جستجو ی کاربر را توسعه داده و باعث می شود نتایج مرتبط بدست آیند

پس روش ترکیبی گسترش کوری چیست ؟

استفاده از ابزار های بروز تر در جهت افزایش کارایی گسترش کوری ! (با توجه به تحقیقات انجام شده)

چه چیزی برای ارایه داریم ؟

می خواهیم در چند سطح با ترکیب کردن روش های موجود و روش جدید ایجاد توسط خودمان کوری را گسترش دهیم .. آیا موفق میشویم ؟ خواهیم دید ؟

معیار کارایی و مقایسه ؟

از دو دیتاست برای آزمایش و ارزیابی داده ها استفاده می کنیم

۱. جستجوی درمیان تویت ها (TREC 2011) و مجموعه ای از متن ها (مقاله های متنشر شده) با موضوعات سخت برای دسته بندی
۲. ساخت توضیح مختصر برای تویت (INEX 2014)

این دیتاست ها خود در آزمایش هایی مورد آزمایش قرار گرفته اند که از نتایج آن ها استفاده خواهیم کرد

صفحه ۴

گسترش کوری به صورت دقیقتر :

اضاف کردن ترم هایی به کوری جستجو شده برای افزایش کارایی !

- مثلاً: نتایج مرتبطی که گوگل به ما نمایش می دهد

روش های گشترش کوری :

Local: ترم هایی را که متن های مرتبط موجود در داکیومنت های برگشت داده شده از اجرای کوری را بر می گرداند

Global: ترم هایی را که از نظر آماری با کوری مرتبط باشند را اضافه میکند

External: ازدانش استخراج شده از منابع خارجی استفاده می کنیم.... ویکی پدیا

روش ترکیبی : می توانیم این روش ها را برای دستیابی به نتایج بهتر باهم ترکیب کنیم!

مشکلات :

یکی از چالش های بازیابی میکروبلاگ ها **عدم تطابق ترم ها به دلیل کوتاهی کوری** می باشد

مشکل گسترش کوری براساس رابطه معنایی بین کوری های مرتبط مواجه خواهیم شد

صفحه ۵

تعاریف اولیه

PRF (pseudo-relevance feedback) بازخورد شبه رابطه ای :

از نتایج کوری ها استفاده کنیم و اینکه این نتایج برای ایجاد یک کوری جدید مرتبط مناسب هستند یا خیر

در این مقاله توضیح می دهیم که چگونه با استفاده از PEF ، منابع خارجی و ترکیب آن با قوانین ارتباط تولید ترم ها و انتخاب آن ها را بهبود میبخشیم ، این عمل مجرب به ایجاد یک روش ترکیبی (hybrid) می شود که ما به آن HQE میگویم

توجه : عملیات ما شامل دو بخش کلی : ۱. ساخت ترم های مناسب ۲. انتخاب ترم های مناسب می باشد

در بخش تولید : از روش استفاده می کنیم

۱. روش آماری که برپایه ی قوانین ارتباط کار میکند که وابستگی قوی را تشخیص دهد
۲. روش معنایی که از جستجوی مقالات ویکی پدیا به بخصوص بخش تعاریف و استخراج داده برای توسعه ی کوری اصلی
۳. روش مفهومی (concept) براساس هستی شناسی دیبی پدیا

{دیبی پدیا : یک سایت که داده های علم شناختی رای با ساختاری خاص (گرافی) دارد و می تواند به ساده کردن کار جستجو کمک کند (شبیه سیستم داخلی گوگل برای جستجو با این تفاوت که در اختیار همه است }

- بهترین عمکرد زمانی بدست می آید که چندین کوری توسعه یافته تولید کنید و بهترین نتایج آن هارا انتخواب کنیم (ممکن است با انتخواب یکی عملکرد ضعیفی داشته باشیم) درنتیجه اگر از یکی از روش ها استفاده نکنیم ممکن است عملکرد ضعیفی داشته باشیم

در بخش انتخاب:

به طور معمول بر اساس میزان تکرار کلمات در متن (doc) صورت میگیرد ، اگر چه تکرار کلمه همواره روش مناسبی برای مشخص کردن ارتباط نیست ، درحالی که برخی کلمات پس زمینه ای هستند

برای غلبه بر مشکل فوق ما از روش دوگانه استفاده میکنیم :

۱. بر قوانین ارتباط بین ترم های کاندید
۲. انتخاب ترم های گسترش مناسب با استفاده از تحلیل معنایی (ESA)

- ما هم چنین عمل گر انتخاب کننده ی ترم ها را با استفاده از یک معیار معنایی جدید (ESAC) که تحلیل صریح معنایی (ESA Explicit Semantic Analysis) ویکی پدیا را با معیار اطمینان قوانین ارتباط ترکیب میکند استفاده میکنیم ، این کار به ما اجازه ی تخمین معنایی بین کوری اصلی و ترم های مرتبط استخراج شده توسط قوانین ارتباط را میدهد
- ESAC هم دانشنامه (ویکی ها) هم وابستگی ترم ها را درنظر میگیرد ، این یک فاکتور کلیدی برای پیداکردن ترم های دقیق میباشد

ESA یک معیار رابطه ی معنایی می باشد که در حوضه ی IR استفاده میشود (تحلیل صریح معنایی)

C : در آخر آن یک معنی خاص برای ما دارد که بخش های بعدی آن را معرفی خواهیم کرد (معیار اطمینان)

صفحه ۶

تعاریف اصلی :

کوری : مجموعه ای تشکیل شده از ترم ها (تی کوچک)

ساپ (تی) : (ساپورت تی) : یک عدد می باشد که تعداد دایکیومنت هایی که شامل تمامی ترم های موجود در مجموعه ی تی بزرگ می باشد را نشان می دهد ، هرچه بزرگ تر باشد احتمالا آن تی مجموعه ی با ارزش تری برای ما خواهد بود

ساپرت نسبی (reletuve support(T)) برابر است با عدد ساپورت تی تقسیم بر تعداد کل متن ها

به مجموعه ی ترم T ترم پرتکرار (frequent) گفته می شود اگر ساپورت آن از آستانه ای که کاربر مشخص می کند بیشتر باشد که این مقدار با minsupp نمایش داده می شود

به یک مجموعه ی ترم بسته (close) گفته میشود اگر هیچ کدام پدران آن (superset) برابر با ترم ست اصلی نباشد

برای قوانین داریم :

ساخت قوانین ارتباط به صورت T1 => T2 (احتمال وجود ترم دوم از حدی بالاتر باشد درحالی که ترم اول در متن وجود داشته باشد ، اگر کلمه ی اول در متن وجود داشته باشد احتمالا دومی هم هست)

- قابل ذکر است که روش های گسترش کوری که بر اساس قوانین ارتباط عمل میکنند نیازی به داشتن دانش قبلی یا پردازش زبانی ندارند! (میزان تاثیر آن های در بازیابی داده در سیستم های آی آر قبلا بررسی شده است)

ضریب اطمینان (Confidence) اگر بیشتر از میزانی باشد به آن قانون معتبر میگوینت و با minconf میگویند

یک حداقل minsup داده میشود تا تمام ترم های پرتکرار متن ساخته شوند

ESA

یک معیار برای رابطه ی معنایی می باشد ، هر موضوع (concept) ویکی پدیا ، که به صورت یک وکتور در آمده این وکتور ها براسا tf * idf ساخته شده اند ، این قدرت بین موضوع و کلمات را دسته بندی میکند

- مقدار سی (c) که در بالا گفته شد اشاره به کانفیدنس (ضریب اطمینان) اشاره دارد

- تحقیقاتی در این حوضه ها انجام شده به صورتی که به آن ها قوانین کتابشناسی یا قوانین داده (information law) گفته میشود ، در این بین ما به قانون Zipf استناد میکنیم! (قوانین ظاهر شدن کلمات پر تکرار) { یک قانین یک انم ، اگر پرتکرار ترین کلمه ان باز ظاهر شده باشد کلمه ی دوم ، ان دوم ظاهر خواهد شد!!!}
- از الگوریتم CHARM برای استخراج قوانین ارتباط استفاده کرده ایم (یک روش اول عمق برای جستجو استفاده میکند)
-

صفحه ۷

مثال : در این صفحه ۳ رابطه تعریف کرده ایم و مقادیر ساپورت و ضریب اطمینان را برای هرکدام حساب کرده ایم .... همان طور که میبینید رابطهی بین کلمات کارخانه و خودرو .....

صفحه ۸

نمای کلی از کل پروسه : در شکل سمت چپ به طور کلی نمای استخراج ترم های کاندید از منابع (دایکیومنت ها و ...) رای میبینید برای ساخت ترم ها و در شکل سمت راست نمای انتخواب ترم ها

۱. شکل چپ

استخراج قوانین ارتباط (با استفاده از روش گلوبال و لوکال)

استخراج ترم ها از منابع دانش خارجی (ویکی و دیبی پدیا)

۲. شکل سمت راست

ترکیب ترم های منتخب از نظر آماری ، معنایی و مفهومی و انتخاب از بین آن ها

صفحه ۹

ساخت ترم های کاندید :

حل مشکل تولید ترم : برای حل آن ما از چندین منبع دانش علاوه بر مجموعه ی متن هایی مثل ویکی پدیا و دیبی پدیا برای متنوع سازی ترم های توسعه استفاده میکنیم

به طور دقیق تر مدل ما ابتدا ترم های مورد نظر را برای توسعه میسازد سپس آن ها را ترکیب میکند (از منابع مختلف) { بر اساس رابطه ی معنایی}

هدف ما افزایش کارایی QE میباشد

بر اساس ترکیب روش های local و global عمل میکند ، فرض ما این است که ترم های مرتبط بیشتر از ترم های نا مرتبط در متن ظاهر می شوند ، از local برای استخراج ترم از متن (corpus) و از general برای استخراج ترم ها براساس قوانین ارتباط ، منابع خارجی ترم های کاندید خود را تولید می کنند

.. توسعه ی آماری

تلاش میکند مجموعه ای از متن ها C را مرتبط با کوری داده شده با استفاده از ابزار استخراج داده بدست آورد

شامل یک تابع محلی (لوکال) همراه با یک تابع کلی (global) میباشد : استفاده PRF و قوانین ارتباط برای انتخاب ترم های مرتبط

با استفاده از قوانین ارتباط وابستگی های قوی بین ترم ها را بدست می آوریم و ترم های کاندید براساس پیشین های قوانین موجود در کوری (q) ساخته می شوند.

.. توسعه ی معنایی

به دانش RS گفته می شود!

- فرض گرفته می شود که متن ها ساختار بندی شده اند! (داده های متنی هستند)

برای این کار از چندین هیورستیک استفاده میکنیم

- همه ی متن های موجد در rs را بر اساس کوری جستجو میکنیم

برای توسعه ی معنایی از دانش ویکی پدیا استفاده میکنیم ، (جستجوی دایکیومنت های موجود در ویکی پیدیا و مرتبط با کوری و جستجو در متن های آنها .....)

..توسعه ی مفهوم

برپایه ی منابع خارجی علم شناختی خارجی ، ترم های مرتبط با مفهموم را اسنخراج کردن (دیبی پدیا)

SPARQL

- مفهوم با استفاده از کلمات هم معنی ، جایگزین و ... ساخته می شود (کلماتی که یک معنی را میرسانند)

صفحه ۱۰

انتخاب ترم های کاندید :

تابع ریلیتد نسس (مرتبط بودن) یک امتیاز برای ترم نسبت به کوری بر میگرداند

ترم با کوری مرتبط در نظر گرفته میشود اگر امتیاز آن از مو بیشتر باشد که در آن مو حداقل استانه باشد

ترم مورد نظر با انتخاب مرتبط ترین ترم صورت میگیرد

از ESAC استفاده میکنیم که ترکیبی از معیار های خطی میباشد

اگر به رابطه ی ارتباط بین ترم و کوری دقت کنید میبینید که یک رابطه ی خطی بین تحلیل معنایی و ضریب اطمینان رابطه ی بدست آماده توسط ما را در نظر می گیرد!!

طوری ESA را نگه میداریم که بیش از حد ضعیف نشود!

(مکس کانفیدنس) بیشترین اطمینان : بین رابطه ی R ، کوری q و ترم t

صفحه ۱۱

تنظیمات HQE: همان طور که در جدول میبینید چندید روش برای عملکرد این روش ارایه شده است :

این روش در دوبخش تولید ترم های کاندید و اتخواب ترم ها ی کاندید دارای چندیدن روش می باشند

استفاده از روش های صرفا آماری ، معنایی ، مفهومی و یا ترکیبی از آن ها ....

صفحه ۱۲

بررسی عملکرد (ولیدیشن)

MAP = mean average (میانگین متوسط ...)

P : تعداد داکیومنت ها ....

BM25

۱۶ ملیون توییت (بدون هیچ داده ی اضاف !) مقایسه با PRF کلاسیک!

تایم استمپ ها را حذف کردیم ....

شامل ۴ بخش شامل ۵۰ تاپیک سخت

با روش های کلاسیک BM25

V fold ...

توییت ها معیار مناسبی نیستند به همین دلیل از ۵۰ هزار داک ویکی پدیا استفاده می کنیم

مینمم ترشهلد (مین کانف) باعث حذف داده های مرزی نا مرتبط می شود ...

آلفا = ۰.۵ و مو (میانگین) = ۰.۴

این جدول برای توییت ها است (معیار اصلی مقایسه با زمان بوده این مدنظر ما نیست ..)

برای هر مدل همه ی مدل ها سیستم ما بهتر عملکرد اما .... (از PRF)

فریم ورک ما که بر دانش خارجی تکیه میکند به شکل واضح سود مند است.

صفحه ۱۳

مقاله های سخت چاچ شده (انگلیسی)

در روش های کلاسیک تر ... TREC 2004

نتایج ما فقط در روش آماری (بدون سلکشن | از بین همه ...) توانستند بهتر از روش های معمول (بیس لاین) عمل کنند ! و همه بدون سلکشن

در ۳۰ دایکیومنت و برای ۵ و ۱۰ نتیجه ی برتر نتایج بهتری تولید میکند

من در تحلیل متن از تیتر + توضیحات و روایت استفاده کرده ایم

نتیجه گیری ما این است (دلیل عملکرد ضعیف تر ...) : نتایج بدست آمده توسط روش آماری به اندازه ی کافی مناسب هستند که فیلتر کردن بخشی از آن ها (ترکیب با سایر روش ها) باعث افت عملکرد آن می شود. (کاهش کیفیت)

همه روش های ما از روش های معرفی شده در این دیتاست بهتر عملکردند با وجود اینکه آن روش ها از ترکیب چهار کوری استفاده می کردند درحالی که روش ما تنها از یک کوری استفاده می کند

با توجه به اینکه این داده ها دایکیومنت هستند و خود به خوبی بخش بندی و ... نوشته شده اند پیدا کردن رابطه های بین کلمات و فیلتر کردن آن ها عملا باعث کاهش عملکرد می شود ولی در توییت ها این چنین نبو د....

این مسیله ها هنوز فاصله ی بسیاری با حل شده بودن دارند ....