مطالعه تطبیقی قطعات شامل پاسخ و قطعات سنتی

Xian-Ling Mao[1], DanWang[1], Yi-Jing Hao[1], Wenqing Yuan[1], and Heyan Huang[1]

[1] Beijing Institute of Technology, Beijing, China
{maoxl,2220150504,hhy63}@bit.edu.cn, wangdan12856@sina.com
2 Beijing Guzhang Mobile Technology Co., Beijing, China
ywq8876@163.com

چکیده. تقریبا هر موتوی جست و جوی متن از قطعات استفاده می کند تا به کاربران در تعیین کردن میزان ارتباط موارد بازیابی شده در فهرست رتبه بندی کمک کند. اگرچه قطعات شامل پاسخ می تواند به طور مستقیم به بهبود کارایی جستجو کمک کند، مطالعه کمی از چنین شهودی دست نخورده باقی ماند. در این مقاله ابتدا یک روش ساده قطعه شامل پاسخ را برای جست وجوی پرسش و پاسخ مبتنی بر جامعه (cQA) پیشنهاد می کنیم، سپس روش خود را با پیشرفته ترین الگوریتم های قطعه سنتی مقایسه می کنیم. نتایج تجربی نشان می دهد که با در نظر گرفتن داوری های ربط و ارزیابی های رضایت بخش اطلاعات، روش قطعه شامل پاسخ، به طور قابل ملاحظه ای بهتر از پیشرفته ترین روش های سنتی اجرا می شود.

کلمات کلیدی

قطعه شامل پاسخ، مطالعه کمی، cQA ، بازیابی اطلاعات

۱ مقدمه

بازیابی اطلاعات، فرآیند فراهم کردن مرتبط ترین اسناد از میان اسناد موجود برای کاربران است [1]. معنای پرس وجو های ارسال شده توسط کاربران به موتور جستجو، هرچند که به عنوان سؤال مشخص شده است، به لحاظ ذاتی مبهم می باشد. برای مثال اگر پرس و جوی وارد شده توسط کاربر "what is the big bang?" باشد، ممکن است منظور کاربر درباره گروه ارکستری با نام "big bang" باشد یا ممکن است درباره "big bang theory" باشد و یا حتی ممکن است نام یک نمایش تلویزیونی به نام "big bang" باشد. مدل های مختلف بازیابی [2-4] تا کنون مطرح شده است. این مدل ها یک فهرست مرتبط از نتایج پرس و جو ها را برمی گردانند. با این حال به دلیل ابهام در معناشناسی پرس و جو، طراحی یک طرح رتبه بندی ثابت که همیشه بتواند میزان ارتباط نتایج پرس و جو را با اهداف کاربر اندازه گیری کند، غیر ممکن است. برای جبران کردن عدم صحت توابع رتبه بندی، قطعه سند تقریبا به یک جزء استاندارد برای موتورهای جستجو به منظور افزایش نتایج پرس و جو تبدیل شده است. یک قطعه برای انعکاس ارتباط بین یک سند و یک پرس و جو استفاده می شد[5,6] که ما آن را "قطعه سنتی" نامیدیم. اگر قطعه ها دارای کیفیت بالایی باشند، کاربران به طور ضمنی می توانند ارتباط نتایج جست وجو را دریابند. و این می تواند به موتور جستجو کمک کند تا ارتباطات درک شده کاربر از اسناد مرتبط را با ربط درست هم تراز کند. بنابراین این مطلب برای موتورهای جست و جو خیلی بحرانی است که قطعه اسنادی تولید کنند که از سوء تفاهم درک ربط اسناد جلوگیری شود[7]. اگر قطعه برای یک سند بسیار مرتبط، ضعیف تولید شده باشد، کاربر ممکن است سند را غیرمرتبط تلقی کند و بر روی آن کلیک نکند. علاوه بر این، کیفیت قطعه نیز یک معیار مهم برای ارزیابی موتورهای جستجو است[8]. به طور سنتی، یک قطعه برای نشان دادن ارتباط بین یک سند و یک پرس و جو استفاده می شده است[5,6]. برای قطعه های سنتی، اساسی ترین و مهم ترین قابلیت، داوری ربط است. این قابلیت قطعه ها، *ربط* نامیده می شود.

به طور مستقیم، اگر یک قطعه حاوی کلماتی باشد که نیاز کاربر را برآورده کند، کاربران از کلیک بر روی سند اصلی اجتناب خواهند کرد و کارایی مرورگر بالاتر می رود. این خصوصیت قطعه ها، *رضایتمندی* نامیده می شود و ما این نوع قطعه ها را *قطعه های شامل پاسخ* می نامیم. هدف اصلی رضایتمندی، بهبود کاربر پسند بودن و کارایی مشابه هدف دست یافتن به یک وظیفه از NTCIR[۱] با یک کلیک است، بدین صورت که نیازاطلاعاتی با اولین خروجی نمایش داده شده سیستم بدون نیاز به کلیک بیشتر، برآورده شود. با این حال، مشخص نیست که کدام یک از قطعات شامل پاسخ و قطعات سنتی بهتر هستند. تا جائیکه ما می

[۱] http://www.thuir.org/1click/ntcir9/

دانیم، هیچ بررسی کمی و کیفی در میان این دو نوع قطعه وجود ندارد. در این مقاله، ما توانایی قطعات شامل پاسخ را در مقایسه با قطعه های سنتی بررسی خواهیم کرد. ما توانایی ربط و رضایتمندی قطعات را در یک بایگانی با مقیاس بزرگ به وسیله جست و جوی سرعت، ارجاعاتی به اسناد پر از متن و بازخورد کاربران ارزیابی خواهیم کرد. سهم کار ما شامل موارد زیر است:

- ما قطعات سنتی و قطعات شامل پاسخ را با هم مقایسه می کنیم.
- ما یک الگوریتم مولد قطعه شامل پاسخ ساده را طراحی می کنیم که می تواند به کاربران کمک کند تا میزان مرتبط بودن را به سرعت داوری کند و هم چنین نیاز اطلاعاتی را به سرعت برآورده کند.
- ما میزان سودمندی قطعات شامل پاسخ را در مقایسه با قطعات سنتی، توسط مطالعات تجربی مان بررسی می کنیم. نتایج امیدبخش این است که مسیر پژوهش ها در آینده، شامل پاسخ هایی در مورد قطعه باشد.

۲ کارمشابه

کارهای اولیه، قطعه های ایستا و مستقل از مستقل پرس و جو تولید کردند که شامل چند جمله اول اسناد بازگشتی می باشد. چنین رویکردی کارآمد است اما اغلب بی نتیجه است[11,12]. انتخاب جملات برای درج در قطعه بر اساس رتبه ای که آنها با کلیدواژه ها مطابقت می کنند، تبدیل به پیشرفته ترین نسخه قطعه متمایل به پرس و جو برای اسناد متنی شده است[5,13,14]. به طور کلی، دو نوع از روش های انتخاب جمله در تولید قطعه مبتنی بر پرس و جو وجود دارد. یکی از این روش ها، روش هیوریستیک است[5,14-17] که در آن اهمیت جملات توسط قوانین هیوریستیکی بیان می شود. به عنوان مثال اگر یک جمله اولین جمله یک پاراگراف باشد، ممکن است اهمیت آن از دیگر جملات بیشتر باشد. روش دیگر، روش یادگیری ماشین است[6,18]. معیارها یا ویژگی ها معمولا در این روش ها استفاده می شود. ویژگی ها عبارتند از اینکه آیا جمله یک عنوان یا خط اول سند است ، تعداد کلمات کلیدی و کلمات کلیدی مجزا در جمله چقدر است و یا تعداد کلمات عنوان که در جمله ظاهر می شود چقدر است و موارد دیگر. با این حال، تکنیک های تولید قطعه که برای اسناد متنی ساده طراحی شده است، قادر به پردازش خوب داده های ساخت یافته و نیمه ساخت یافته نیستند و هم چنین نمی توانند اطلاعات ساختاری داده ها را اهرم بندی کنند و بنابراین عملکرد خوبی ندارند.

اخیرا قطعه های ساختاری برای اسناد XML مورد بررسی قرار گرفته اند[7].Huang et al . [7] مجموعه ای از نیازمندی ها را برای قطعه ها پیشنهاد می دهند و یک الگوریتم جدید را طراحی می کنند که به صورت موثر قطعه های کوچک اما حاوی اطلاعات مفید را تولید می کنند. [19].Ellkvist et al چگونگی تولید قطعه برای داده های گردش کار را با در نظر گرفتن

اطلاعات ساختاری دانه ریز، مورد بررسی قرار داده است. با این حال، تا جاییکه ما اطلاع داریم، هنوز هیچ روش تولید قطعه برای داده های cQA وجود ندارد، این داده ها هنوز هم نسخهٔ قطعه متمایل به پرس و جو را به طور کلی در نظر گرفته اند و همان طور که از مطالعات کاربر مشاهده می شود، عملکرد خوبی ندارند.

۳ تنظیمات آزمایشی

در cQA، همه می توانند در هر موضوعی سوال بپرسند و جواب دهند، و کسانی که در جستجوی اطلاعات هستند به کسانی که پاسخ سوالات را می دانند، متصل می شوند. به دلیل این که معمولا پاسخ ها به صراحت توسط انسان ارائه شده اند، می تواند در پاسخ دادن به سوالات دنیای واقعی نیز مفید باشد[9]. اگرچه دستیابی به *رضایتمندی* برای جست وجوی عمومی دشوار است اما برای جستجوی CQA نسبتا ساده تر است، زیرا یک سوال از یک سند CQA دارای پاسخ های متناظر می باشد. جملات این پاسخ ها را می توانند برای برآوردن نیازهای اطلاعاتی کاربر مورد استفاده قرار دهند. در ضمن، هدف این مقاله، تأیید استفاده از قطعه های شامل پاسخ است ، نه استخراج پاسخ از اسناد. بنابراین در این مقاله، بایگانی داده های cQA را به عنوان مجموعه داده آزمایشی انتخاب خواهیم کرد، و برای رسیدن به اهداف پروهشی مان بر روی آن ها تمرکز خواهیم کرد.

۳–۱ مجموعه داده

ما برای ساخت یک مجموعه داده جامع برای آزمایشات مان، تقریبا تمام جفت های سوال و پاسخ (جفت های QA) را از دو دسته برتر و مشهور (کامپیوتر و اینترنت، بهداشت و سلامت) که شامل چهل و سه زیرمجموعه از داده های Yahoo!Answer[۲] از تاریخ سپتامبر ۲۰۰۵ تا سپتامبر ۲۰۰۸ است، جمع آوری کرده ایم. منبع داده ذکر شده یک بایگانی شامل ۶،۳۴۵،۷۸۶ جفت QA را فراهم می کند. هر cQA مربوط به سند شامل *عنوان سوال*، *بدنه سوال*، *بهتری پاسخ*، *پاسخ های دیگر* و *توضیحات* می باشد.

۳–۲ قطعه های شامل پاسخ

برای طراحی یک روش مناسب قطعه شامل پاسخ برای جستجوی CQA، ما باید دو سوال را در نظر بگیریم: (۱) چگونه پاسخ ها را رتبه بندی کنیم ؟ (۲) چه کنیم اگر حجم یک پاسخ زیاد باشد؟

---

[۲] http://answers.yahoo.com/.

- **اعتبار پاسخ:** ما کیفیت پاسخ را از سه منظر بررسی می کنیم. اولا، مردم می توانند بهترین پاسخ انتخاب شده توسط شخص سوال کننده یا سیستم QA در یک سند cQA را به عنوان باکیفیت ترین پاسخ در نظر بگیرند. دوما با استفاده از تعداد رأی که بازتاب محبوبیت پاسخ هاست، پاسخی با بیش ترین تعداد رأی می تواند به عنوان باکیفیت ترین پاسخ در میان پاسخ ها باشد. نهایتا، پاسخی که توسط بالاترین مقام ارشد ارسال می شود، می تواند به عنوان باکیفیت ترین پاسخ در نظر گرفته شود.

- **اعتبار بهترین پاسخ:**

$$BestAnsImp(ans) = \mathbf{1}_{IsBestAns}(Type(ans)) \times (\alpha \times \mathbf{1}_{ChosedBySys}(BestAnsType(ans)) + (1-\alpha) \times \mathbf{1}_{ChosedByAsker}(BestAnsType(ans))) \quad (1)$$

فرمول (1) یک تابع شاخص است؛ تابع Type(ans) نوع پاسخ را بر می گرداند، نوع پاسخ می تواند IsBestAns و یا NotBestAns باشد. تابع BestAnsType(ans) نوع بهترین پاسخ ans را بر می گرداند که می تواند ChosedBySys یا ChoosedByAsker باشد. α نشان دهنده ضریب اطمینان برای بهترین پاسخ انتخاب شده توسط سیستم است.

- **اعتبارتعداد رأی:**

$$VoteImp(ans_i, doc) = \frac{VoteNum(ans_i)}{\sum_{ans_j \in doc} VoteNum(ans_j)} \quad (2)$$

طبق فرمول بالا، اهمیت امتیاز i امین پاسخ ansi در سند cQA به نام doc ، توسط تناسب تعداد رأی ansi اندازه گیری می شود.

- **اعتبار صلاحیت پاسخ دهنده:** ما به سادگی می توانیم امتیاز اعتبار صلاحیت i امین کاربر $user_i$ را توسط فرمول زیر به دست بیاوریم:

$$AuthImp(user_i) = \frac{BestAnsNum(user_i)}{AllAnsNum(user_i)} \qquad (3)$$

در فرمول ذکر شده، BestAnsNum(*useri*) به تعداد بهترین پاسخ های ارسال شده توسط کاربر$user_i$ اشاره دارد و هم چنین AllAnsNum(*useri*) نشان دهنده تمام پاسخ های ارسال شده توسط کاربر $user_i$ است .

در پیاده سازی مان، تمام این سه عامل در نظر گرفته می شود تا کیفیت پاسخ را اندازه گیری کنند. ما به سادگی، تمامی این سه عامل را به وسیله ترکیب خطی شان به صورت زیر با هم ترکیب کردیم:

$$AnswerImp(ans_i, doc) = \alpha BestAnsImp(ans_i) + \beta VoteImp(ans_i, doc) + \gamma AuthImp(user_i) \qquad (4)$$

در این فرمول α، β و γ نشان دهنده وزن این سه عامل هستند؛ $\alpha+\beta+\gamma=1$. پاسخی با بالاترین امتیاز محاسبه شده توسط فرمول ۴، به عنوان باکیفیت ترین پاسخ در نظر گرفته می شود. در تمامی آزمایشات مان، برای α مقدار ۰,۶، برای β مقدار ۰,۳ و برای γ مقدار ۰,۱ در نظر گرفته شده است.

- **ملاحظات اندازه:**

ما ابتدا توزیع تعداد کلمه در هر بخش (بدنه پرسش، بهترین پاسخ و پاسخ ها) را از اسناد cQA در مجموعه داده مان به دست آوردیم. ما متوجه شدیم که توزیعها قوانین قدرت را دنبال می کنند.

روابط قانون قدرت نشان می دهد که اندازه بسیاری از مولفه ها کوچک است، فقط تعداد کمی از مولفه ها دارای اندازه بزرگ هستند. در اینجا مولفه به یکی از سه قسمت پرسش، بهترین پاسخ و پاسخ ها اطلاق می شود. نتایج نشان داده شده در جدول ۱ کمی هستند.

جدول ۱. آمار های مربوط به طول (بر حسب تعداد کلمات)بدنه پرسش(QBody)، بهترین پاسخ(BAns) و تمام پاسخ ها(AAns)

Words	<30	<50	<100	<150	<250	<350
QBody	0.5011	0.6819	0.8935	0.9584	0.9964	0.9986
BAns	0.3653	0.5476	0.7908	0.8883	0.9565	0.9776
AAns	0.5191	0.6965	0.8841	0.9445	0.9804	0.9902

بنابراین ما می توانیم حد آستانه T را برای فیلتر کردن پاسخ ها انتخاب کنیم. اگر تعداد کلمات در هر پاسخ بیشتر از T باشد، ما از T کلمه برای نمایش دادن پاسخ ها استفاده می کنیم. و گرنه پاسخ را به عنوان باکیفیت ترین پاسخ در نظر می گیریم.

بنابراین، ما یک چارچوب قطعه شامل پاسخ را برای جستجوی cQA که شامل سه بخش عنوان، بدنه سوال و باکیفیت ترین پاسخ است، طراحی کرده ایم. به طور خلاصه، الگوریتم مطرح شده ابتدا سند cQA را برای به دست آوردن تمام قسمت ها، از جمله موضوع سوالات، پرسش و پاسخ ها، تجزیه می کند. سپس الگوریتم تمام پاسخ ها را برای به دست آوردن باکیفیت ترین پاسخ توسط فرمول ۴، رتبه بندی می کند. اگر تعداد کلمات باکیفیت ترین پاسخ بیشتر از حد آستانه T باشد الگوریتم فقط T کلمه را به عنوان باکیفیت ترین پاسخ ذخیره می کند. سوم، الگوریتم تمام اطلاعات بیش از حد و کمتر قابل توجه را از بدنه سوال حذف می کند و یک بدنه سوال تمیز به دست می آورد. در نهایت، الگوریتم بدنه پرسش تمیز و باکیفیت ترین پاسخ را به عنوان قطعه CQA از سند cQA باز می گرداند.

۳-۳ الگوریتم قطعه پایه

پیشرفته ترین روش تولید قطعه که توسط Metzler et al.[18] مطرح شده است، به عنوان الگوریتم پایه ای ما انتخاب شده است. این الگوریتم از روش یادگیری گرادیان درخت تقویت شده تصمیم گیری برای فاز تولید قطعه استفاده می کند. ویژگی های پذیرفته شدهT عبارت دقیق پرس و جو، نسبت همپوشانی، نسبت همپوشانی هم زمان ، مدل زبانی جمله ، طول و محل جمله است. ۱۰ پرس و جوی نمونه برداری شده از بین سوالات در مجموعه داده های مان و ۲۰ سند بازیابی شده مربوط به آنها، به عنوان داده های آموزشی مورد استفاده قرار گرفت. از یک ارزیابی کننده انسانی خواسته شد که تمامی ۲۰۰ صفحه را با استخراج جملات بر اساس پرس وجوهای متناظر خلاصه کند. تعداد پیشنهاد شده برای جملات یک خلاصه، ۵ جمله است. با این حال، اگر

تعداد کمتر یا بیشتری از جملات متناسب وجود داشته باشد، بر خلاف پیشنهاد داده شده می تواند انتخاب شود. برای GBDT ها، ما از بسته GBM برای R[۳] استفاده می کنیم.

۳-۴ مدل بازیابی

برای بازیابی کردن، ما از یک تابع رتبه بندی مشابه تابعی که توسط Xue et al. [4]، مطرح شده است، استفاده می کنیم، که براساس کار قبلی مبتنی بر مدل های بازیابی مبتنی بر ترجمه ساخته می شود و تلاش می کند تا بر برخی از نقاط ضعف خود غلبه کند. این تابع به صورت زیر فرمول بندی می شود:

$$P(\mathbf{q}|(q,a)) = \prod_{w \in \mathbf{q}} P(w|(q,a)) \quad (5)$$

$$P(w|(q,a)) = (1-\lambda)P_{mx}(w|(q,a)) + \lambda P_{ml}(w|C) \quad (6)$$

$$P_{mx}(w|(q,a)) = \alpha P_{ml}(w|q) + \beta \sum_{t \in q} P(w|t)P_{ml}(t|q) + \gamma P_{ml}(w|a) \quad (7)$$

در فرمول های بالا، q همان سوال[۴] کاربر است و C به کل مجموعه بایگانی اشاره دارد، $C=\{(q,a)_1,(q,a)_2,\ldots,(q,a)_L\}$. λ پارامتر هموارسازی برای مجموعه C است. و $P_{ml}(w|C) = \frac{\#(w,c)}{|C|}$ تخمینی برای حداکثر احتمال است در حالیکه |C| نشان دهنده اندازه مجموعه C است. $P(w|t)$ احتمال ترجمه شدن عبارت سوالی t به عبارت پرس و جوی w است. ما اهمیت ربط را توسط پارامتر های α،β و γ کنترل می کنیم و $\alpha+\beta+\gamma=1$.

یکی از تفاوت ها به نسبت مدل اصلی Xue et al. [4] این است که در معادله ۶، ما از هموارسازی Jelinek-Mercer به جای هموارسازی Dirichlet که توسط Delphine et al. [10] انجام شده است، استفاده می کنیم. تفاوت دیگر این است که ما از مدل ترجمه آماری کلمه آموزش داده شده توسط Delphine et al. [10] استفاده می کنیم، که عملکرد بهتری به نسبت مدل اصلی Xue et al دارد. در تمامی آزمایشات مان، برای α مقدار ۰,۵ ، برای β مقدار ۰,۳ و برای λ مقدار ۵.۰ در نظر گرفته شده است.

---

[۳] http://cran.r-project.org/web/packages/gbm/.

[۴] در این مقاله، منظور از سوال کاربر همان پرس و جوی کاربر است.

۳–۵ معیارهای ارزیابی

ما دو رویه تجربی داریم: رویه تجربی برای داوری ربط و رویه تجربی برای رضایتمندی اطلاعات. به دلیل کار مشابه، معیارهای مطرح شده توسط Tombros et al. [5] در این بخش پذیرفته خواهد شد. معیار های پذیرفته شده برای آزمایش، موارد زیر هستند:

a) معیار فراخوانی، دقت و F_1 مربوط به داوری ربط.

b) سرعت اجرای داوری.

c) نیاز به ارزیابی کنندگان برای دریافت کمک از متن کامل اسناد بازیابی شده.

d) دیدگاه ذهنی کاربران در مورد کمک ارائه شده توسط قطعه هر سند بازیابی شده.

معیار فراخوانی، دقت و F_1: این معیارها اغلب برای ارزیابی کردن داوری های ربط استفاده می شود که به صورت زیر محاسبه می شود:

$$P = \frac{N_{cr}}{N_{tir}}; R = \frac{N_{cr}}{N_{tr}}; F_1 = \frac{2 \times P \times R}{P + R} \qquad (8)$$

در این فرمول، N_{cr} نشان دهنده تعداد اسناد مرتبط برای یک پرس و جو که به درستی توسط ارزیابی کننده تشخیص داده شده است؛ N_{tir} نشان دهنده تعداد کل اسناد مرتبط در میان موارد مورد بررسی برای پرس و جوی مورد نظر است؛ N_{tr} تعداد کل اسناد مرتبط نشان داده شده برای پرس و جوی موردنظر است.

۵۰ پرس و جو (سوال) مورد استفاده در دو رویه تجربی از سوالات موجود در مجموعه داده مان به عنوان نمونه انتخاب شدند؛ در عین حال، ۳۰ سند برتر بازیابی شده برای هر پرس و جو به عنوان پایه و اساس کارمان، به صورت دستی مورد داوری ربط قرار می گیرند.

۴ تجزیه و تحلیل تجربی

۴–۱ آزمایشات ربط

در این بخش، ما عملکرد کاربران را در فرایند داوری ربط بین اسناد بازیابی شده و پرس وجوهای خاص (به عنوان مثال سوالات) بررسی می کنیم. این فرآیند شامل دو کار برای داوری ربط اسناد موجود در فهرست رتبه بندی شده، چه با الگوریتم قطعه پایه و یا الگوریتم قطعه شامل پاسخ است. برای دستیابی به این هدف، دو گروه که هر یک شامل ۱۰ ارزیاب بودند، دعوت شدند. ارزیاب ها به صورت تصادفی به یک گروه نسبت داده شدند، و هر گروه مسئول انجام فقط یک کار است[5]. برای هر پرس و جو، ارزیاب ها توسط پرس و جو و یک لیست اسناد بازیابی شده با قطعه ارائه شدند و گفته شد که لیست، نتایج بازیابی شده برای یک پرس و جوی خاص است. تنها اقداماتی که ارزیاب ها می توانستند انجام دهند، حرکت کردن در طول لیست یا کلیک بر روی متن کامل سند cQA بود. بنابراین هدف آنان این بود که در ۲ دقیقه، حداکثر اسناد مرتبط را تشخیص دهند. نتایج حاصل از رویه تجربی در قسمت ذیل، ارائه شده و مورد تحلیل قرار گرفته است.

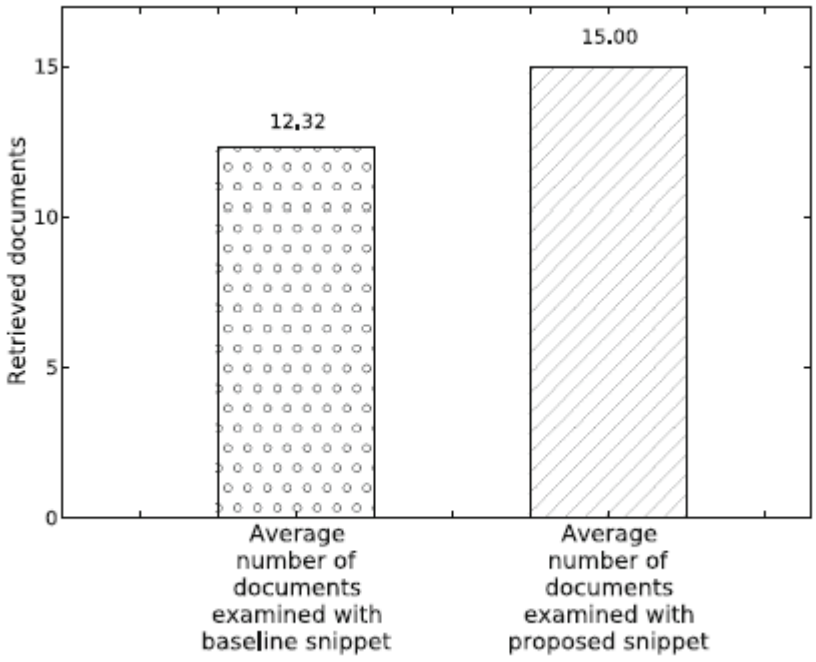
فراخوانی، دقت و F_1 : هم چنان که در جدول [۵] ۳ مشاهده می کنیم؛ مقادیر صحت، فراخوانی و F1 برای گروه ارزیابی که از الگوریتم قطعه پیشنهاد شده استفاده می کنند به طور قابل ملاحظه ای بیشتر از مقادیر مربوط به گروه ارزیابی هستند که از الگوریتم قطعه پایه استفاده کردند: تفاوت کارایی ۲۰,۲۵٪ ، ۶,۳٪ و ۱۱,۳۶٪ است. ما نتیجه می گیریم ارزیاب هایی که از الگوریتم قطعه پیشنهاد شده برای بازیابی فهرست اسناد cQA استفاده کردند، در داوری های ربط عملکرد بهتری نسبت به ارزیاب هایی که از پیشرفته ترین الگوریتم سنتی قطعه استفاده کردند، داشتند. بنابراین این نشان دهنده این است که الگوریتم قطعه پیشنهاد شده به کاربران اجازه می دهد که اسناد cQA مرتبط بیشتری را شناسایی کنند و آن ها را دقیق تر شناسایی کنند.

جدول۲. مقادیر P ، R و F_1 مربوط به دو گروه ارزیاب

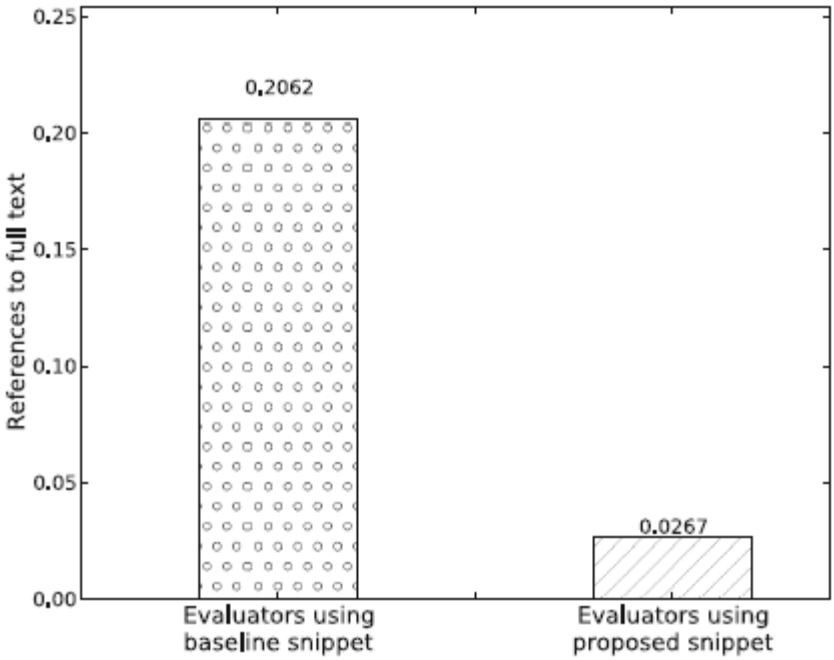
	Precision	Recall	F_1
Baseline	0.5944	0.4676	0.5234
Proposed	**0.7969**	**0.5306**	**0.6370**

[۵] نتایج نشان داده شده در جدول ۲ با میانگین گرفتن نتایج به ازای هر پرس و جو بر روی تعداد کل پرس و جو ها به دست آمده است، بنابراین مقادیر میانگین فراخوانی، دقت و F1 به ازای هر پرس و جو تولید شده است.

سرعت. سرعت نتایج در شکل ۱ نشان داده شده است. ما میانگین تعداد اسنادی که از الگوریتم قطعه پایه و الگوریتم قطعه پیشنهاد شده استفاده کردند، بررسی کردیم. شکل نشان می دهد که ارزیاب هایی که از الگوریتم قطعه پیشنهاد شده استفاده کردند، به طور متوسط برای هر پرس و جو تعداد ۱۵ سند را برگردانده اند، در حالیکه متوسط تعداد سند برگردانده شده توسط ارزیاب های دیگر ۱۲,۳۲ است. این مقدار به میزان ۲۱,۷۵٪ افزایش در میانگین تعداد اسناد مورد بررسی داشته است. بنابراین، یک تمایل صریحی وجود دارد برای کاربرانی که از الگوریتم قطعه پیشنهاد شده استفاده کردند، این که در داوری های ربط، عمکلرد سریع تری نسبت به کاربرانی داشته باشند که از الگوریتم قطعه پایه استفاده کردند.



شکل ۱. نتایج سرعت

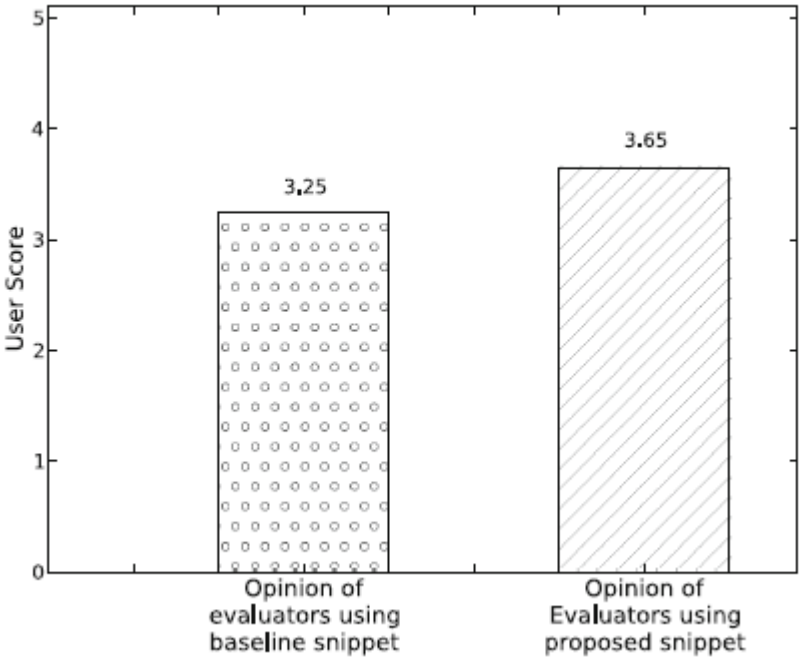


شکل۲. میانگین تعداد ارجاعات به متن کامل اسناد(به ازای هر پرس و جو)

ارجاع به متن کامل اسناد. داده های جمع آوری شده از ارجاعات کاربران به متن کامل اسناد نشان می دهد ارزیاب هایی که از الگوریتم قطعه پایه استفاده کردند، مجبور به ارجاع دادن به ۲,۵۴ متن کامل اسناد به ازای هر پرس و جو شدند، در حالیکه ارزیاب های متعلق به گروه آزمایشی دیگر به طور متوسط به ۰,۴ متن کامل اسناد مراجعه کردند. اگر ما این مقادیر را با میانگین تعداد اسنادی که هر گروه آزمایشی به ازای هر پرس و جو بررسی کردند نرمال کنیم، نتایج نشان داده شده در شکل ۲ به دست می آید. متن کامل ۲۰,۶۲٪ از اسناد به ازای هر پرس و جو نیاز به ارجاع توسط ارزیاب هایی که از الگوریتم قطعه پایه استفاده می کنند، دارد. حال آن که این میزان برای ارزیاب هایی که از الگوریتم قطعه شامل پاسخ استفاده می کنند، ۲,۶۷٪ است.

این تفاوت می تواند به وضوح به اطلاعات قطعه ای که همراه هر سند بازیابی شده است، مربوط شود. ما در میابیم که روش پیشنهاشده عملکرد بهتری دارد. کاربران برای تعیین میزان مرتبط بودن اسناد، نیاز به سرنخ کمتری دارند، و به خصوص آن ها به سرنخ هایی در مورد زمینه ای که نوع سوال پرس و جو از آن تولید می شود، نیاز دارند. علاوه بر این، نتایج ما نشان می دهد قطعه پیشنهاد شده شواهد کافی برای ارزیاب ها ارائه می کند تا از داوری های ربط شان پشتیبانی کنند.

دیدگاه کاربران. به عنوان یک مهر تایید نتایج حاصل از مقوله های قبلی، دیدگاه ذهنی کاربران، از پرسشنامه ای که از آنها خواسته شد تا بعد از جلسه خود آن را تکمیل کنند، جمع آوری شد. طبق این دیدگاه، میزان استفاده از قطعه پیشنهادی بالاتر از قطعه پایه ارزیابی شد. این نتیجه در شکل ۳ نشان داده شده است که در آن محدوده مقیاس، از ۱ (کم ترین میزان سودمندی) تا ۵ (بیش ترین میزان سودمندی) است. داده ها نشان می دهد ارزیاب هایی که از قطعه پایه استفاده کردند، امتیاز ۳,۲۵ را به متوسط اطلاعات همراه دادند ، حال آن که امتیاز نشان داده شده توسط ارزیاب هایی که به مورد دیگر اختصاص داده شده بوند، ۳,۶۵ است. این مطلب نشان دهنده این است که کاربران به شواهد بیش تری برای میزان مرتبط بودن اسناد بازیابی شده نیاز دارند و قطعات شامل پاسخ روی همین نیازمندی تمرکز کرده اند.



شکل۳. دیدگاه ذهنی ارزیابی کنندگان

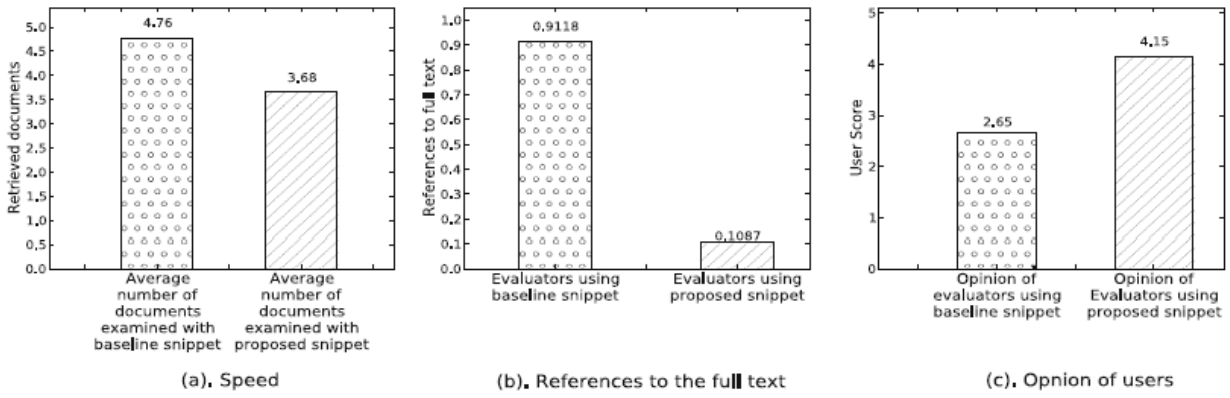
۴-۲ آزمایشات رضایتمندی

در اینجا، ما دو کار را بررسی می کنیم: برای دریافت کلماتی که بتواند نیاز اطلاعاتی یک فهرست رتبه بندی شده را چه از روش قطعه پایه و یا قطعه شامل پاسخ برآورده کند. ارزیاب ها به صورت تصادفی به یک گروه نسبت داده شدند، و هر گروه مسئول انجام فقط یک کار است. برای هر پرس و جو، ارزیاب ها توسط پرس و جو و یک لیست اسناد بازیابی شده با قطعه ارائه شدند و گفته شد

که لیست، نتایج بازیابی شده برای یک پرس و جوی خاص است. تنها اقداماتی که ارزیاب ها می توانستند انجام دهند، حرکت کردن در طول لیست یا کلیک بر روی متن کامل سند cQA بود. بنابراین، هدف آنها دستیابی به اطلاعاتی بود که پرس و جو را برآورده کند و تا زمانیکه اطلاعات به دست بیاید، صبر کند.

نتایج به دست آمده در طول رویه آزمایشی، در شکل ۴ نمایش داده شده است.

سرعت. شکل۴.(a) نشان می دهد ارزیاب هایی که از قطعه پیشنهاد شده استفاده کردند، به میانگین ۳,۶۸ سند به ازای هر پرس و جو برای برآورده کردن نیاز اطلاعاتی دست یافتند، در حالی که ارزیاب هایی که از قطعه پایه استفاده کردند، به میانگین ۴,۷۶ سند برای آوردن نیاز اطلاعاتی دست یافتند. اگرچه این تفاوت کم است اما به میزان ۲۲,۶۹ درصد کاهش در تعداد اسناد مورد بررسی است. بنابراین، ما می توانیم نتیجه بگیریم که یک تمایل صریحی وجود دارد برای کاربرانی که از الگوریتم قطعه شامل پاسخ استفاده کردند، این که نیاز اطلاعاتی را به نسبت کاربرانی که از الگوریتم قطعه پایه استفاده کردند، سریع تر برآورده کنند.



شکل ۴. سرعت، ارجاعات به متن کامل اسناد و دیدگاه های کاربران برای تابع رضایتمندی

ارجاع به متن کامل اسناد. ارزیاب هایی که از الگوریتم قطعه پایه استفاده کردند، مجبور به ارجاع دادن به ۴,۳۴ متن کامل اسناد به ازای هر پرس و جو شدند، در حالیکه ارزیاب های متعلق به گروه آزمایشی دیگر به طور متوسط به ۰,۴ متن کامل اسناد مراجعه کردند. اگر ما این مقادیر را با میانگین تعداد اسنادی که هر گروه آزمایشی به ازای هر پرس و جو بررسی کردند نرمال کنیم، نتایج نشان داده شده در شکل ۴.(b) به دست می آید. متن کامل ۹۱,۱۸٪ از اسناد به ازای هر پرس و جو نیاز به ارجاع توسط ارزیاب هایی که از الگوریتم قطعه پایه استفاده می کنند، دارد . حال آن که این میزان برای ارزیاب هایی که از الگوریتم قطعه شامل پاسخ استفاده می کنند، ۱۰,۸۷٪ است.

این تفاوت می تواند به وضوح به اطلاعات قطعه ای که همراه هر سند بازیابی شده است، مربوط شود. نتایج پیش فرض اولیه را تایید می کند که روش قطعه پیشنهاد شده، می تواند عملکرد بهتری داشته باشد برای این که به کاربران دربرآوردن نیاز اطلاعاتی کمک کند. اگر اطلاعات کاربران نمی تواند به وسیله قطعه برآورده شود، کاربران به متن کامل اسناد مراجعه می کنند. نتایج ما نشان می دهد قطعه پیشنهادشده، شواهد کافی برای برآوردن نیاز اطلاعاتی در اختیار ارزیاب ها قرار می دهد.

دیدگاه کاربران. در شکل ۴.(c) محدوده مقیاس، از۱ (کم ترین میزان سودمندی) تا ۵ (بیش ترین میزان سودمندی) است. داده های نشان داده شده در این شکل حاکی از آن است ارزیاب هایی که از قطعه پایه استفاده کردند، امتیاز ۲,۶۵ را به متوسط اطلاعات همراه دادند ، حال آن که امتیاز نشان داده شده توسط ارزیاب هایی که به مورد دیگر اختصاص داده شده بوند، ۴,۱۵ است.

در طی بحث های پس از آزمایش، کاربرانی که از قطعه پایه استفاده کردند، نارضایتی خود را نسبت به اطلاعاتی که به آن ها ارائه شده بود ابراز کردند. به طور خاص، آنها بر این واقعیت تأکید کردند که تقریبا برای هر سندی که می خواستند میزان برآورده کردن نیاز اطلاعاتی را بررسی کنند، مجبور بودند به متن کامل آن مراجعه کنند. از این رو، نتیجه بحث های پس از آزمایش در عین حال دلیل دیگری بر خوب بودن فرضیه ساخته شده است، که کاربران نیاز به کلماتی دارند که بتوانند نیاز اطلاعاتی را که در قطعه وجود دارد، برآورده کنند. قطعه های شامل پاسخ که دارای پاسخ با کیفیت بالا باشند، تمرکز بیش تری بر برآورده کردن این نیازمندی دارند.

۵ نتایج و کارهای آینده

تا جایی که ما اطلاع داریم، این اولین کاری است که به مسئله تولید قطعه های شامل پاسخ برای جستجو cQA می- پردازد. در همین حال، مطالعات کمی ما نشان می دهد روش قطعه شامل پاسخ که به طور قابل ملاحظه ای به لحاظ داوری های ربط و ارزیابی های رضایتمندی اطلاعات از پیشرفته ترین روش های سنتی بهتر عمل می کند، نشان دهنده جهت تحقیقات امیدبخش به سمت قطعه شامل پاسخ در آینده است.

سپاس گذاری. این برنامه توسط ۸۶۳ برنامه (2015AA015404)، بنیاد ملی علوم چین(61402036،60973083 ، 61273363)، فناوری پروژه پکن(Z151100001615029)، پروژه برنامه ریزی علم و فناوری ایالت گوانگدونگ(2015A020217002 ,2014A010103009)، پروژه برنامه ریزی علوم و فناوری گوانگژو(201604020179).

منابع

1. Campos, R., Dias, G., Jorge, A.M., Jatowt, A.: Survey of temporal information retrieval and related applications. ACM Comput. Surv. **47**(2), 1–41 (2015)
2. Jeon, J., Croft, W.B., Lee, J.H.: Finding similar questions in large question and answer archives. In: ACM International Conference on Information and Knowledge Management, pp. 84–90. ACM (2005)
3. Lee, J.T., Kim, S.B., Song, Y.I., Rim, H.C.: Bridging lexical gaps between queries and questions on large online qa collections with compact translation models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, 25–27 October 2008, Honolulu, A Meeting of Sigdat, A Special Interest Group of the ACL, pp. 410–418 (2008)
4. Xue, X., Jeon, J., Croft, W.B.: Retrieval models for question and answer archives. In: International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR, Singapore, pp. 475–482, July 2008
5. Tombros, A., Sanderson, M.: Advantages of query biased summaries in information retrieval. In: Proceedings of ACM SIGIR, pp. 2–10 (1998)
6. Wang, C., Jing, F., Zhang, L., Zhang, H.J.: Learning query-biased web page summarization. In: Sixteenth ACM Conference on Information and Knowledge Management, CIKM, Lisbon, pp. 555–562, November 2007
7. Huang, Y., Liu, Z., Chen, Y.: Query biased snippet generation in XML search. In: ACM SIGMOD International Conference on Management of Data, pp. 315–326. ACM (2008)
8. He, J., Shu, B., Li, X., Yan, H.: Effective time ratio: a measure for web search engines with document snippets. In: Cheng, P.-J., Kan, M.-Y., Lam, W., Nakov, P. (eds.) AIRS 2010. LNCS, vol. 6458, pp. 73–84. Springer, Heidelberg (2010)
9. Zhou, G., Zhou, Y., He, T., Wu, W.: Learning semantic representation with neural networks for community question answering retrieval. Knowl. Based Syst. **93**, 75–83 (2015)
10. Bernhard, D., Gurevych, I.: Combining lexical semantic resources with question and answer archives for translation-based answer finding. In: ACL 2009, Proceedings of

the, Meeting of the Association for Computational Linguistics and the, International Joint Conference on Natural Language Processing of the AFNLP, 2–7 August 2009, Singapore, pp. 728–736 (2009)

11. Edmundson, H.P.: New methods in automatic extracting. J. ACM **16**(2), 264–285 (1969)
12. Gomez-Nieto, E., San, R.F., Pagliosa, P., Casaca, W., Helou, E.S., Oliveira, M.C.,et al.: Similarity preserving snippet-based visualization of web search results. IEEE Trans. Vis. Comput. Graph. **20**(3), 457–470 (2014)
13. Silber, H.G., Mccoy, K.F.: Efficiently computed lexical chains as an intermediate representation for automatic text summarization. Comput. Linguist. **28**(4), 487–496 (2002)
14. Turpin, A., Tsegay, Y., Hawking, D.,Williams, H.E.: Fast generation of result snippets in web search. In:SIGIR 2007: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, pp. 127–134, July 2007
15. Goldstein, J., Kantrowitz, M., Mittal, V., Carbonell, J.: Summarizing text documents:sentence selection and evaluation metrics. In: Research and Development in Information Retrieval, pp. 121–128 (1999)
16. Joho, H., Hannah, D., Jose, J.M.: Emulating query-biased summaries using document titles. In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 709–710. ACM (2008)
17. Ichikawa, K., Morishita, S.: A simple but powerful heuristic method for accelerating k-means clustering of large-scale data in life science. IEEE/ACM Trans. Comput. Biol. Bioinf. (TCBB) **11**(4), 681–692 (2014)
18. Metzler, D.: Machine learned sentence selection strategies for query-biased summarization. In: SIGIR Learning to Rank Workshop (2008)
19. Ellkvist, T., Strmbck, L., Lins, L.D., Freire, J.: A first study on strategies for generating workflow snippets. In: International Workshop on Keyword Search on Structured Data, pp. 15–20(2009)