بازیابی اطلاعات

دکتر امین گلزاری اسکوئی

a.golzari@azaruniv.ac.ir

a.golzari@tabrizu.ac.ir

https://github.com/Amin-Golzari-Oskouei

دانشگاه صنعتی ارومیه

پاییز ۱۴۰۲

فصل ۲

مجموعه واژگان، عبارات و لیست پست‌ها

مطالب این فصل

- درک واحد پایه سیستم‌های کلاسیک بازیابی اطلاعات: واژه‌ها و اسناد
- سند چیست؟ اصطلاح چیست؟
- چگونه از متن خام به کلمات (یا نشانه‌ها) برسیم؟

اسناد

- آخرین موضوع: سیستم بازیابی بولی ساده
  - فرضیات ما این بود:
    1. ما می‌دانیم که سند چیست.
    2. ما می‌توانیم هر سند را به صورت ماشینی بخوانیم.
- این می‌تواند در واقعیت پیچیده باشد.

تجزیه سند (Parsing a Document)

- ما باید به فرمت و زبان هر سند بپردازیم.
- متن ما در چه قالبی قرار دارد؟ (html , word , excel , pdf , …)
- متن ما به چه زبانی می‌باشد؟
- چه مجموعه حروفی در آن استفاده شده است؟
- هر یک از اینها یک مسئله طبقه بندی است که در ادامه این دوره به بررسی آن خواهیم پرداخت.
- پیشنهاد دیگر: استفاده از روش‌های فراابتکاری است.

پیچیدگی (فرمت یا زبان)

- یک فهرست واحد معمولا شامل اصطلاحات چندین زبان است.
  - گاهی اوقات یک سند یا اجزای آن حاوی چندین زبان یا فرمت است. (ایمیل فرانسوی با پیوست اسپانیایی)
  - واحد سند برای نمایه سازی چیست؟
- انتخاب واحد سند برای شاخص گذاری چیست؟ (یک فایل؟، یک ایمیل؟ یک ایمیل با 5 پیوست؟، گروهی از فایل‌ها (ppt یا latex در html)؟)
- در نتیجه ؛ پاسخ به سوال "سند چیست؟" بی اهمیت نیست و نیاز به برخی تصمیمات طراحی دارد.

تعاریف اصطلاحات

- کلمه (word): یک رشته محدود از کاراکترها همانطور که در متن ظاهر می‌شود.
- اصطلاح (term): یک کلمه "نرمال" (تاریخ تحولات لغوی، هجی(آوا حروف کلمات) و...)
- Token: نمونه ای از یک کلمه یا اصطلاح که در یک سند وجود دارد.
- Type: در بیشتر موارد مانند یک اصطلاح است، یک کلاس معادل (هم‌ارز) از نشانه‌ها.

نرمالسازی (دسته کردن هم ارزی عبارات)

- با شکستن اسناد (و همچنین پرس و جو) به نشانه‌ها، ساده ترین حالت این است که آیا نشانه‌ها در پرس و جو با نشانه‌ها در لیست نشانه‌های سند تطبیق می‌کند؟ هر چند، موارد بسیاری وجود دارد که دو دنباله کاراکتر کاملا یکسان نیستند اما شما تمایل دارید که تطبیقی رخ دهد.
- برای نمونه، اگر شما به دنبال USA هستید، ممکن است امیدوار باشید اسنادی را که شامل U.S.A. هستند نیز بیابید. نرمانسازی نشانه‌ها روند استاندارد سازی نشانه‌ها است به طوریکه تطبیق، علیرغم تفاوت‌های صوری در دنباله کاراکتری نشانه‌ها، رخ دهد. متداولترین روش برای نرمالسازی این است که دسته‌های هم ارزی به طور ضمنی ایجاد کنیم که به طور مرسوم بعد از یک عضو مجموعه نامگذاری می‌شوند. برای نمونه، اگر نشانه‌های anti-discriminatory و antidiscriminatory هر دو به عبارت antidiscriminatory، هم در متن سند و هم در پرس و جوها، نگاشت شوند، جستجو برای یک عبارت، اسنادی که شامل دیگری است را نیز بازیابی می‌کند.

- ما معمولا به طور ضمنی کلاس‌های هم ارزی اصطلاحات را تعریف می‌کنیم.
- روش دیگر انبساط نامتقارن را انجام دهید:

Enter	Search
window	window , windows
windows	Windows , windows , Windows (no expansion)
Windows	Windows

- قدرتمندتر، اما ناکارآمدتر
- چرا نمی‌خواهید window، Window، Windows، windows را در یک کلاس معادل سازی قرار دهید؟
- نرمال سازی و language detection interact

PETER WILL NICHT MIT. → MIT = mit

He got his PhD from MIT. → MIT ≠ mit

- نرمالسازی در بسیاری از موارد مفید به نظر می‌رسد اما می‌تواند زیان بخش نیز باشد. در حقیقت شما می‌توانید نگران بسیاری از جزئیات دسته کردن هم ارزی باشید، اما اغلب به این نتیجه می‌رسید که اگر پردازش به طور پایدار برای پرس و جو و اسناد انجام شود، جزئیات ممکن است زیاد روی عملکرد تاثیری نداشته باشد.

- در برخی زبان‌ها اعراب باعث تغییر معنی یک کلمه می‌شود، در این موارد که ممکن است نرمالسازی زیان بخش باشد بهترین راه چیست؟ در این موارد بهترین راه معادل گرفتن تمامی کلمات با صورت بدون اعراب آنها است.

- در متن‌ها و جستجوها نوشتن حروف به صورت بزرگ و کوچک مسئله دیگری است که باید حل شود. یک استراتژی رایج برای غیر حساس کردن به حروف کوچک و بزرگ، تبدیل تمامی حروف به حروف کوچک است.

یادآوری: ساختار شاخص معکوس

- Input:

Friends, Romans, countrymen.	So let it be with Caesar	. . .

- Output:

friend	roman	countryman	so	. . .

- هر توکن کاندیدای ورودی پست است.

- توکن‌های معتبر برای انتشار چیست؟

تمرین

In June, the dog likes to chase the cat in the barn. – How many word tokens? How many word types?

چرا نشانه گذاری یا Tokenize دشوار است؟

– even in English. Tokenize: *Mr. O'Neill thinks that the boys' stories about Chile's capital aren't amusing.*

مشکلات نشانه گذاری (Tokenization Problem)

One word or two ? (or several ?)

- Hewlett-Packard
- State-of-the-art
- co-education
- the hold-him-back-and-drag-him-away maneuver
- data base
- San Francisco
- Los Angeles-based company
- cheap San Francisco-Los Angeles fares
- York University vs. New York University

Numbers

- 3/20/91
- 20/3/91
- Mar 20, 1991
- B-52
- 100.2.86.144
- (800) 234-2333
- 800.234.2333
- Older IR systems may not index numbers...
- ... but generally it’s a useful feature.

莎拉波娃现在居住在美国东南部的佛罗里达。今年4月9日，莎拉波娃在美国第一大城市纽约度过了18岁生日。生日派对上，莎拉波娃露出了甜美的微笑。

- هر زبان جدید مسائل جدیدی ایجاد خواهد کرد، برای مثال زبان چینی: صورت قطعه‌قطعه نشده متن زبان چینی با استفاده از کاراکترهای ساده شده زبان چینی.
- هیچ فاصله ی خالی بین کلمات وجود ندارد، حتی بین جملات نیز وجود ندارد – فاصله‌ظاهری بعد از نقطه چینی ( ) تنها یک خطای چاپی است که به دلیل قرار گرفتن کاراکتر در سمت چپ کادر مربعی آن است. جمله اول تنها کلمات با کاراکتر چینی بدون فاصله بین آنهاست.
- جملات دوم و سوم شامل ارقام عربی و علامت گذاری است که کاراکترهای چینی را قطع کرده است.

和尚

- ابهام در قطعه بندی کلمات زبان چینی.
- دو کاراکتر را می‌توان به عنوان یک کلمه به معنای "راهب" در نظر گرفت و یا به عنوان دنباله‌ای از دو کلمه با معنی "وَ" (حرف ربط) و "هنوز".

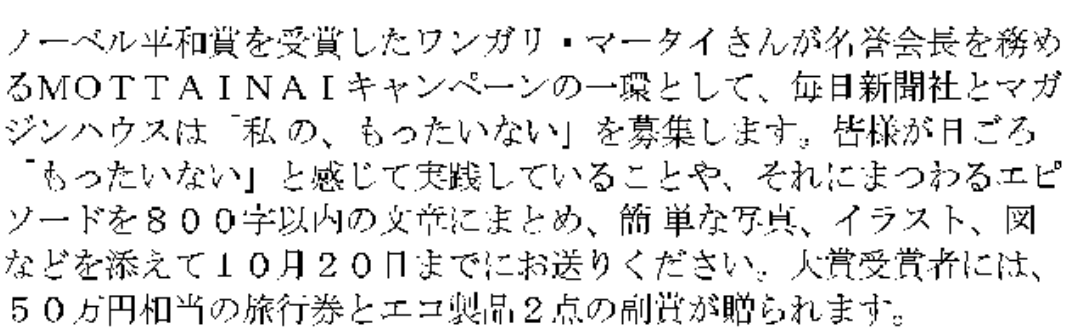
❖ دیگر نمونه‌های بدون فاصله (whitespace)

برای مثال زبان آلمانی کلمات مرکب را بدون فاصله می‌نویسد مانند:

Computerlinguistik → Computer + Linguistik

Lebensversicherungsgesellschaftsangestellter →leben+versicherung+gesellschaft+ angestellter

❖ سیستم‌های بازیابی برای زبان آلمانی، از روال **جداکننده مرکب** استفاده می‌کنند، که معمولا با مشاهده اینکه یک کلمه می‌تواند به کلمات متعدد که در مجموعه واژگان حضور دارند، تقسیم شود. این پدیده به سر حد محدودیات در زبان‌های آسیا شرقی می‌رسد.


ノーベル平和賞を受賞したワンガリ・マータイさんが名誉会長を務めるMOTTAINAIキャンペーンの一環として、毎日新聞社とマガジンハウスは「私の、もったいない」を募集します。皆様が日ごろ「もったいない」と感じて実践していることや、それにまつわるエピソードを８００字以内の文章にまとめ、簡単な写真、イラスト、図などを添えて１０月２０日までにお送りください。大賞受賞者には、５０万円相当の旅行券とエコ製品２点の副賞が贈られます。

- زبان ژاپنی، زبانی با سیستم نگارش بسیار دشوار است که در شکل بالا نمونه ای از آن را می‌بینیم. زبان ژاپنی مدرن به طور استاندارد ترکیبی از چندین الفبا، کاراکترهای زبان چینی، دو هجایی (hiragana ، katakana) و کاراکترهای غربی (حروف لاتین، اعداد عربی و نشانه‌های مختلف) است.
- اگر چه قوانین استاندارد سازی و پایدار زیادی در طول دوره تحصیلی برای انتخاب سیستم نگارش وجود دارد، در بسیاری از موارد یک کلمه با سیستم‌های نگارشی مختلف نوشته می‌شود. به عنوان مثال، یک کلمه را می‌توان به صورت katakana برای تاکید نوشت یا گاهی به صورت hiragana، یا گاهی با کاراکتر زبان چینی نوشت. بازیابی موفق نیازمند دسته کردن هم ارزی پیچیده ای برای سیستم نگارشی است.

ك ِ ت ا ب ٌ ⇐ كِتَابٌ

un b ā t i k

/kitābun/ *‘a book’*

❖ مثالی از کلمه عربی استاندارد تلفظ شده.

استقلت الجزائر في سنة 1962 بعد 132 عاما من الاحتلال الفرنسي.

← START ← → ← →

‘Algeria achieved its independence in 1962 after 132 years of French occupation.’

- ترتیب خطی مفهومی از کاراکترها لوزما آن ترتیبی نیست که شما روی صفحه می‌بینید. زبان‌هایی که از راست به چپ نوشته می‌شوند، مانند عبری و عربی، می‌تواند متن پراکنده ی از چپ به راست نیز داشته باشد، مانند اعداد یا قیمت دلار. با داشتن مفاهیم مدرن نمایش Unicode، ترتیب کاراکترها در فایل‌ها با ترتیب مفهومی آنها تطبیق می‌یابد و معکوس کاراکترهای نمایش داده شده با سیستم تحویل اداره می‌شود، اما این ممکن است برای اسناد با کد گذاری‌های قدیمی‌تر صحت نداشته باشد.

- résumé vs. resume (حذف ساده لهجه)
- Umlauts: Universität vs. Universitaet (جایگزینی با دنباله حروف “ae”)

- مهمترین معیار: کاربران چگونه می‌توانند پرس و جوهای خود را برای این کلمات بنویسند؟
- حتی در زبان‌هایی که به طور استاندارد دارای لهجه هستند، کاربران اغلب آنها را تایپ نمی‌کنند. (لهستانی؟)

Case Folding

- همه حروف را به حروف کوچک تبدیل کنید.
- استثناهای احتمالی: کلمات بزرگ در وسط جمله

  - ✓ MIT vs. mit
  - ✓ Fed vs. fed

- اغلب بهترین کار این است که همه چیز را کوچک کنید زیرا کاربران بدون توجه به حروف بزرگ از حروف کوچک استفاده می‌کنند.

Stop Words

❖ Stop Words: کلمات بسیار رایج که به نظر می‌رسد ارزش کمی برای کمک به انتخاب اسناد مطابق با نیاز کاربر دارند.

✓ Examples: *a، an، and، are، as، at، be، by، for، from، has، he، in، is، it، its، of، on، that، the، to، was، were، will، with*

❖ Stop Word (توقف) در سیستم‌های (IR) قدیمی‌تر استاندارد بود.

❖ اما برای عبارت‌ها به کلمات توقف نیاز دارید،

✓ Example: “King of Denmark”

اصطلاحات

الگوریتم پورتر

- رایج ترین الگوریتم برای ریشه یابی انگلیسی
- نتایج نشان می‌دهد که حداقل به خوبی ریشه یابی شده است
- 5 مرحله
- مراحل به صورت متوالی اعمال می‌شوند.
- این الگوریتم دارای یک‌سری قوانینی است که با استفاده از آنها پیشوند یا پسوند کلمات را حذف می‌کند.:
  - مثال: "Ement" نهایی را با توجه به باقیمانده‌ای که بیشتر از 1 حرف باشد را پاک کنید

✓ replacement → replac

✓ cement → cement

- از میان قوانین یک دستور مرکب، پسوندی را انتخاب کنید که برای طولانی ترین پسوند اعمال می‌شود.

Lemmatization

اصطلاحات

ریشه یابی یا مدخل‌گیری

- تعریف مدخل‌گیری: فرآیند ابتکاری که انتهای کلمات را برای دستیابی واژه سازی اصولی با دانش زبانی زیادی انجام می‌دهد، قطع می‌کند.
- وابسته به زبان
- اغلب عطفی و اشتقاقی
- مثال برای اشتقاقی:

✓ *automate, automatic, automation* all reduce to *automat*

Lemmatization

اصطلاحات

❖ فرم‌های عطفی/متغیر را به شکل پایه تبدیل کنید:

- ✓ Example: *am, are, is* → *be*
- ✓ Example: *car, cars, car's, cars'* → *car*
- ✓ Example: *the boy's cars are different colors* → *the boy car be different color*
- ✓ Lemmatization implies doing "proper" reduction to dictionary headword form (the lemma).
- ✓ Inflectional morphology (*cutting* → *cut*) vs. derivational morphology (*destruction* → *destroy*)

اصطلاحات

هم ارزی بیشتر

- ✓ Soundex: IIR 3 (phonetic equivalence، Muller = Mueller)
- ✓ Thesauri: IIR 9 (semantic equivalence، car = automobile)

مثال‌هایی از الگوریتم ریشه یابی پورتر

مثال‌ها	قوانین
caresses → caress	SSES → SS
ponies → poni	IES → I
caress → caress	SS → SS
cats → cat	S →

آیا ریشه یابی موثر می‌باشد؟

- ❖ به طور کلی ریشه یابی باعث افزایش اثربخشی برای برخی از پرس و جوها شده و گاها سبب کاهش اثربخشی برای برخی دیگرمی‌شود.

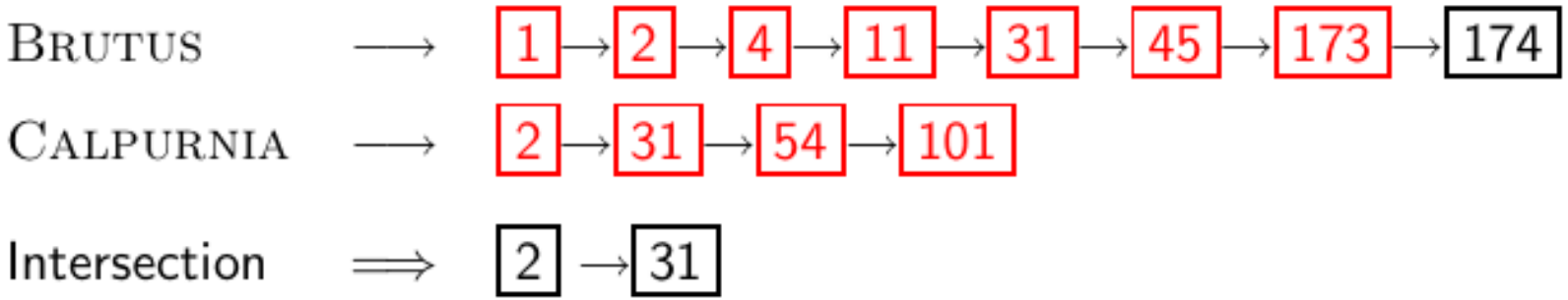
- ✓ Queries where stemming is likely to help: [tartan sweaters], [sightseeing tour san francisco] (equivalence classes: {sweater,sweaters}, {tour,tours})
- ✓ Porter Stemmer equivalence class *oper* contains all of *operate operating operates operation operative operatives operational.*
- ✓ Queries where stemming hurts: [operational AND research], [operating AND system], [operative AND dentistry]

Exercise: What does Google do?

- ✓ Stop words
- ✓ Normalization
- ✓ Tokenization
- ✓ Lowercasing
- ✓ Stemming
- ✓ Non-latin alphabets
- ✓ Umlauts
- ✓ Compounds
- ✓ Numbers

Skip Pointers

Recall basic intersection algorithm

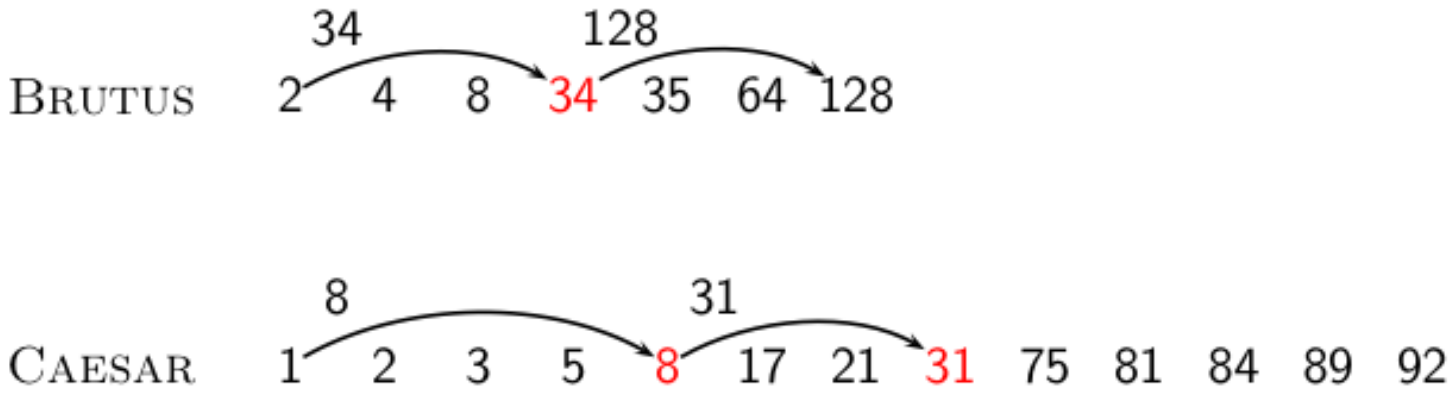


- ✓ Linear in the length of the postings lists.
- ✓ Can we do better?

Skip Pointers

- نادیده گرفتن اشاره گرها به ما امکان می‌دهند از پست‌هایی که در نتایج جستجو ظاهر نمی‌شوند بگذریم.
- باعث افزایش بهره وری الگوریتم Intersection می‌شود.
- کی می‌توانیم اشاره گرها را نادیده بگیریم
- چگونه از بهره وری الگوریتم Intersection مطلع شویم؟

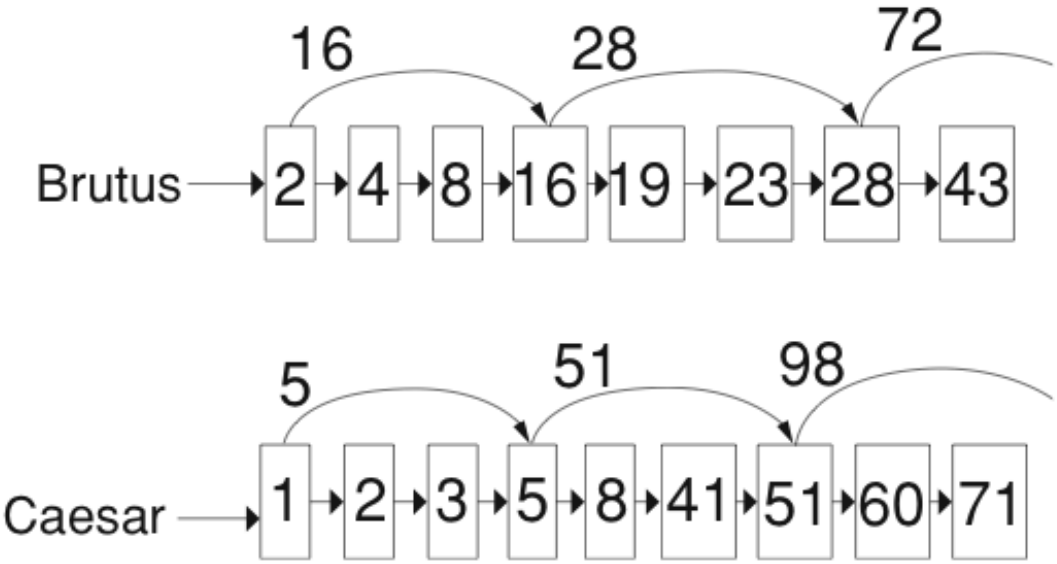
Skip Pointers



فرض کنید در لیست‌های شکل بالا تا جایی پیش می‌رویم که 8 را تطبیق داده و به لیست نتایج انتقال دهیم. هر دو اشاره‌گر را پیش می بریم که 41 در لیست بالایی و 11 در لیست پایینی است. به جای اینکه به سادگی اشاره‌گر پایینی را جلو ببریم، ابتدا اشاره‌گر پرش لیست را بررسی می‌کنیم و متوجه می‌شویم که 31 نیز از41 کوچکتر است بنابراین می‌توانیم اشاره گر پرش لیست را دنبال کنیم.

Skip Pointers

Skip lists: Larger example



Skip Pointers

Intersection with skip pointers

```
INTERSECTWITHSKIPS(p1, p2)
 1  answer ← ⟨ ⟩
 2  while p1 ≠ NIL and p2 ≠ NIL
 3  do if docID(p1) = docID(p2)
 4        then ADD(answer, docID(p1))
 5             p1 ← next(p1)
 6             p2 ← next(p2)
 7        else if docID(p1) < docID(p2)
 8                then if hasSkip(p1) and (docID(skip(p1)) ≤ docID(p2))
 9                        then while hasSkip(p1) and (docID(skip(p1)) ≤ docID(p2))
10                                do p1 ← skip(p1)
11                        else p1 ← next(p1)
12                else if hasSkip(p2) and (docID(skip(p2)) ≤ docID(p1))
13                        then while hasSkip(p2) and (docID(skip(p2)) ≤ docID(p1))
14                                do p2 ← skip(p2)
15                        else p2 ← next(p2)
16  return answer
```

Skip Pointers

پرش‌ها را کجا قرار دهیم؟

یک مصالحه وجود دارد، پرش‌های بیشتر یعنی فواصل پرش کوتاهتر و یعنی اینکه احتمال پرش بیشتری داریم. اما مقایسات بیشتری با اشاره گرهای پرش و فضای ذخیره سازی بیشتری برای پرش نیاز است.
پرش‌های کمتر به معنای تعداد مقایسات کمتر است اما فواصل طولانی پرش به معنای فرصت کمتری برای پرش است. یک روش ساده برای قراردادن پرش‌ها، که در عمل به خوبی نتیجه داده است، این است که برای لیست پست‌ها به طول P، از فاصله ی مساوی $\sqrt{p}$ برای اشاره گرهای پرش استفاده کنیم. روش مطرح شده می‌تواند بهبود یابد.

Skip Pointers

پست‌های موقعیتی و پرس و جوی‌های اصطلاح

- ما میخواهیم به پرسشی مانند [Stanford University] به عنوان یک اصطلاح پاسخ دهیم.
- جمله ای در یک سند مانند The inventor Stanford Ovshinsky never went to university یک تطبیق نباشد. بیشتر موتورهای جستجوی اخیر گیومه نقل قول مستقیم (“Stanford University”) را برای پرس و جوی اصطلاح پشتیبانی میکنند و نشان داده شده است که به راحتی فهمیده میشود و توسط کاربران به طرز موفقیت آمیز به کار می‌رود.
- حدود 10 درصد از پرس و جوهای وب عبارتی (اصطلاحی)هستند.
- دیگر استفاده از لیست‌هایی از اسناد که عبارات واحد را شامل میشوند کافی نیست.
- دو روش برای گسترش شاخص وارونه:

✓ biword index

✓ positional index

Skip Pointers

Biword indexes(شاخص‌های دو کلمه ای)

- هر جفت عبارت متوالی در متن را به عنوان یک عبارت نمایه کنید.

- For example, *Friends*, *Romans*, *Countrymen* would generate two biwords: *"friends romans"* and *"romans countrymen"*

- اکنون می‌توان به عبارات دو کلمه ای به راحتی پاسخ داد.

Skip Pointers

Longer phrase queries

- در این مدل، هر یک از این دو کلمه ای‌ها را به عنوان یک عبارت مجموعه واژگان در نظر میگیریم. پردازش پرس و جوهای دو کلمه عملی و سریع است. اصطلاحات طولانی تر میتوانند یا کوتاه تر شدن، پردازش شوند. پرس و جوی Stanford university palo alto میتواند به پرس و جوی بولی دو کلمه ای شکسته شود.

✓ “Stanford university” AND “university palo” AND “palo alto”

- این پرس و جو میتواند در عمل به خوبی کار کند اما ممکن است مثبت‌های کاذب را نیز نتیجه دهد.
- بدون آزمایش اسناد، نمیتوانیم تایید کنیم که اسنادی که با پرس و جوی بولی ذکر شده تطابق دارند، واقعا شامل اصطلاح 4 کلمه ای اصلی هستند.

- Parse each document and perform part-of-speech tagging
- Bucket the terms into (say) nouns (N) and articles/prepositions (X)
- Now deem any string of terms of the form NX*N to be an *extended biword*
- Examples: catcher in the rye

  N X X N

  king of Denmark

  N X N
- Include extended biwords in the term vocabulary
- Queries are processed accordingly

مشکل شاخص دو کلمه ای (byword indexes)

- چرا شاخص دو کلمه ای به ندرت استفاده می‌شود؟

همانطور که در بالا اشاره شد ؛ False Positives

index blowup به دلیل اصطلاحات طولانی لغت نامه

شاخص‌های موقعیتی (Positional indexes)

- شاخص‌های موقعیتی جایگزین کارآمدتری نسبت به شاخص‌های دو کلمه ای هستند.
- پست لیست‌ها در شاخص‌های غیر موقعیتی: هر پست فقط یک شناسه مدرک (docID) است.
- پست لیست‌ها در شاخص‌های موقعیتی: هر پست یک شناسه مدرک و لیستی از موقعیت‌ها است.

مثال شاخص‌های موقعیتی

Query: *“to*$_1$ *be*$_2$ *or*$_3$ *not*$_4$ *to*$_5$ *be*$_6$*”* TO, 993427:

‹ 1: ‹7, 18, 33, 72, 86, 231›;

2: ‹1, 17, 74, 222, 255›;

4: ‹8, 16, 190, 429, 433›;

5: ‹363, 367›;

7: ‹13, 23, 191›;... ›

BE, 178239:

‹ 1: ‹17, 25›;

4: ‹17, 191, 291, 430, 434›;

5: ‹14, 19, 101›;... › Document 4 is a match!

Proximity search

- دیدیم که چگونه از شاخص‌های موقعیتی برای جستجو اصطلاح استفاده کنیم.
- همچنین برای Proximity search هم می‌توانیم استفاده کنیم.
- برای مثال: employment /4 place ؛ همه اسنادی که شامل EMPLOYMENT و PLACE در فاصله 4 کلمه از هم هستند را پیدا کن.

جمله:

Employment agencies that place healthcare workers are seeing growth is a hit.

Employment agencies that have learned to adapt now place healthcare workers is not a hit.

Proximity search

- Use the positional index
- Simplest algorithm: look at cross-product of positions of (i) EMPLOYMENT in document and (ii) PLACE in document
- Very inefficient for frequent words, especially stop words
- Note that we want to return the actual matching positions, not just a list of documents.
- This is important for dynamic summaries etc.

❖ الگوریتمی برای اشتراک مجاورت لیست پست‌های P1 , P2.

الگوریتم مکان‌هایی را که دو عبارت درون K کلمه ظاهر می‌شوند، می‌یابد و لیستی از سه تایی‌ها شامل شناسه سند و موقعیت عبارت در P1 و P2 را برمی‌گرداند.

```
POSITIONALINTERSECT(p1, p2, k)
 1  answer ← ⟨ ⟩
 2  while p1 ≠ NIL and p2 ≠ NIL
 3  do if docID(p1) = docID(p2)
 4        then l ← ⟨ ⟩
 5             pp1 ← positions(p1)
 6             pp2 ← positions(p2)
 7             while pp1 ≠ NIL
 8             do while pp2 ≠ NIL
 9                do if |pos(pp1) − pos(pp2)| ≤ k
10                      then ADD(l, pos(pp2))
11                      else if pos(pp2) > pos(pp1)
12                                then break
13                   pp2 ← next(pp2)
14                while l ≠ ⟨ ⟩ and |l[0] − pos(pp1)| > k
15                do DELETE(l[0])
16                for each ps ∈ l
17                do ADD(answer, ⟨docID(p1), pos(pp1), ps⟩)
18                pp1 ← next(pp1)
19             p1 ← next(p1)
20             p2 ← next(p2)
21        else if docID(p1) < docID(p2)
22                then p1 ← next(p1)
23                else p2 ← next(p2)
24  return answer
```

طرح‌های ترکیبی

- شاخص‌های دو کلمه ای و موقعیتی می‌توانند به خوبی ترکیب شوند.
- بسیاری از کلمات دوگانه بسیار متداول هستند مانند: Michael Jackson

- For these biwords, increased speed compared to positional postings intersection is substantial.
- Combination scheme: Include frequent biwords as vocabulary terms in the index. Do all other phrases by positional intersection.

- Williams و دیگران (2004) طرح پیچیده تری که هر دو نوع شاخص را با شاخص کلمه بعد به عنوان یک روش میان دو استراتژی اول به کار می‌برد، ارزیابی کرده است.برای هر عبارت، شاخص کلمه بعد عباراتی را که بعد از آن در سند رخ می‌دهند، ثبت می‌کند. آنها نتیجه گرفتند که این استراتژی اجازه می‌دهد که ترکیب متعارف پرس و جوهای عبارات وب، در یک چهارم زمان استفاده از شاخص موقعیتی، انجام شود، در حالیکه 26% فضای بیشتر نسبت به استفاده از شاخص موقعیتی مصرف می‌کند.

کوئری‌های موقعیتی در گوگل

- کوئری‌های موقعیتی بسیار پر هزینه تر از کوئری‌های بولی معمولی برای موتور جستجو وب هستند.
- مثالی از phrase queries: “Stanford University”
- چرا پرس و جوهای موقعیتی از پرس و جوهای بولی معمولی پر هزینه تر هستند؟
- آیا می‌توانید در گوگل نشان دهید که پرس و جوهای اصطلاحی پر هزینه تر از پرس و جوهای بولی هستند؟

تشکر

سوال؟

a.golzari@azaruniv.ac.ir

a.golzari@tabrizu.ac.ir

https://github.com/Amin-Golzari-Oskouei