بررسی کاربردهای نمایش طیفی کلمات در گسترش پرس‌وجو

دانشجو: ارسلان سالاری

استاد راهنما: دکتر آزاده شاکری

دانشکده مهندسی برق و کامپیوتر، دانشگاه تهران

مقدمه

بازیابی اطلاعات به علم پیدا کردن اسناد مرتبط با نیاز اطلاعاتی یک کاربر از میان حجم انبوهی از اسناد خام گفته می‌شود. کاربردهای این علم را می‌توان در موتورهای جست‌وجو، سیستم‌های پیشنهاد دهنده و کاتالوگ کتابخانه‌ها دید . سیستم‌های بازیابی اطلاعات نیاز اطلاعاتی کاربر را در قالب یک پرس‌وجو مشتمل بر چند واژه دریافت می‌کنند و اسناد موجود در بانک اسناد را با توجه به میزان ارتباط با پرس‌وجو مرتب می‌کنند و بر می‌گردانند.

روش‌هایی برای بهبود نتایج معرفی شده‌اند که با گسترش پرس‌وجوی اولیه با استفاده از بازخورد مستقیم یا غیرمستقیم کاربر پرس‌وجوی جدیدی ایجاد می‌کنند و سعی می‌کنند نتایج اولیه را با پرس‌وجوی جدید بهبود ببخشند. یکی از معیارهایی که می‌تواند در امتیازدهی اسناد مفید باشد، استفاده از ارتباط معنایی اسناد با پرس‌وجو است. می‌توان از ارتباط معنایی لغات برای گسترش پرس‌وجو نیز استفاده کرد.

مدل پیشنهادی

در این تحقیق سعی داریم با گسترش عبارت پرس‌وجو در مسئله بازیابی اطلاعات با استفاده از الگوریتم kl-divergence[2] سعی کنیم نتایج نهایی را بهبود ببخشیم.

ما در این مسئله اسناد به دست آمده از نتایج pseudo relevance feedback را به عنوان اسناد مرتبط در نظر می‌گیریم. اینکه لغات موجود در اسناد مرتبط را چطور و با چه احتمالی به عبارت پرس‌وجو نزدیک کنیم اهمیت زیادی دارد و این موضوع اصلی این تحقیق را تشکیل می‌دهد. برای گسترش عبارت پرس‌وجو ما ابتدا تمام لغات موجود در اسناد بازخورد را در نظر می‌گیریم. موضوعی که باید درنظر گرفت این است که همه این لغات برای بهتر شدن بازیابی مفید نیستند بلکه برخی بازیابی را بدتر کرده و برخی تاثیری در نتیجه نمی گذارند.

به همین دلیل هر لغت را به یکی از سه دسته: لغت خوب، لغت بد، و لغت خنثی تقسیم می کنیم. پس از دسته بندی لغات موجود در اسناد بازخورد سعی داریم که با طراحی یک شبکه عصبی یاد بگیریم که هر لغت داده شده برای هر عبارت پرس‌وجو در کدام کلاس قرار دارد. و در نهایت با توجه به جواب شبکه عصبی (یک احتمال تعلق به کلاس لغات خوب) تصمیم میگیریم که لغت را به عبارت پرس‌وجو اضافه کنیم یا نه.

ساختار شبکه عصبی پیشنهادی:

ما تصمیم گرفتیم که از یک شبکه عصبی چند لایه با یک لایه مخفی استفاده کنیم که به عنوان ورودی از تفاوت نمایش طیفی کلمه(glove[1]) گسترش و میانگین نمایش طیفی کلمات موجود در عبارت پرس‌وجو استفاده میکند.

$$input = Exp - \frac{\sum_{q_i \in Q} qi}{|Q|}$$

$$h_1 = relu(W_1 \times input + b_1)$$

$$probs = softmax(W_2 \times h_1 + b_2)$$

شکل(۱) ساختار شبکه عصبی پیشنهادی

نتایج

دیتاست مربوطه که در این تحقیق استفاده شده است دیتاست AP است که شامل مقالات با کیفیت اخبار است که اطلاعات دقیق آن در جدول زیر قابل مشاهده است:

جدول(۱) اطلاعات دقیق دیتاست AP

ID	collection	Queries(title only)	#docs
AP	Associated Press 88-89	TREC 1-3 Ad-Hoc Track, topics 51-200	165k

همچنین نتایج بدست آمده از تحقیق ما نسبت به نتایج پایه پیاده سازی شده برای این دیتاست در جدول زیر قابل مشاهده است.

جدول(۲) مقایسه نتایج تحقیق با روش های پایه

Dataset	Metric	MLE	AWE	Our method
AP	MAP	۰.۲۳۱۲	۰.۲۳۰۴	**۰.۲۴۰۲**

جمع بندی

با توجه به اینکه بازیابی اطلاعات بسیار پرکاربرد است، (همانطور که همه ما هرروزه از موتورهای جست‌وجوی تحت وب استفاده می‌کنیم) می‌توان گفت که بهبود بخشیدن به نتایج موجود هرچند جزئی می‌تواند بسیار مفید واقع شود.

مراجع اصلی

[1]Pennington, Jeffrey, Richard Socher, and Christopher Manning. "Glove: Global vectors for word representation." Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014.

[2] J. Lafferty and C. Zhai. Document Language Models, Query Models, and Risk Minimization for Information Retrieval. In SIGIR ’01, pages 111–119, 2001.