بسمه تعالی

دانشگاه تهران

پردیس دانشکده های فنی

دانشکده مهندسی برق و کامپیوتر

پیشنهاد و فرم حمایت از پایان‌نامه تحصیلات تکمیلی

کارشناسی ارشد ■ **دکتری** □

شماره مرجع * :

* شماره مرجع، توسط معاونت پژوهشی پردیس دانشکده‌های فنّی هنگام صدور ابلاغ درج خواهد شد.

۱- خلاصه اطلاعات پایان نامه

عنوان پایان نامه به زبان فارسی:

تحلیل نظری بازیابی اطلاعات بین زبانی

عنوان پایان نامه به زبان انگلیسی:

Theoretical analysis of cross language information retrieval

نوع پایان نامه: ■ **بنیادی** ■ **کاربردی** □ **توسعه‌ای**

پردیس/دانشکده: فنی **دانشکده/گروه: برق و کامپیوتر**

مقطع تحصیلی : کارشناسی ارشد **رشته و گرایش تحصیلی: مهندسی‌کامپیوتر-نرم‌افزار**

تاریخ پیشنهاد **تاریخ تصویب:**

۲- اطلاعات اساتید راهنما و مشاورین

نوع مسئولیت	نام و نام‌خانوادگی	مرتبه علمی	محل خدمت	امضاء
استاد راهنما (مجری)	**دکتر آزاده شاکری**	استادیار	دانشگاه تهران	
استاد راهنمای دوم (حسب نیاز)				
استاد مشاور				
استاد مشاور دوم (برای دکتری)				

۳- اطلاعات دانشجو

نام و نام‌خانوادگی: علی منتظرالقائم **شماره دانشجویی:** ۸۱۰۱۹۳۲۸۴

رشته و گرایش تحصیلی: مهندسی‌کامپیوتر-نرم‌افزار **دانشکده:** برق و کامپیوتر **مقطع تحصیلی:** کارشناسی ارشد

پست الکترونیک :ali.montazer@ut.ac.ir **تلفن ثابت:** **تلفن همراه:** ۰۹۳۷۴۳۸۸۶۲۰

۴- مشخصات موضوعی پایان نامه

تعریف مسأله، هدف و ضرورت اجرای (حداکثر سه صفحه)

جستجو در بین تعداد زیادی سند یکی از مسائل سنتی در حوزه بازیابی اطلاعات و داده کاوی است که در دهه‌ی ۸۰ میلادی اکثر فعالیت‌های تحقیقاتی حوزه بازیابی اطلاعات را به خود اختصاص داده است. حال آنکه با مرور زمان، سهم اسناد غیر انگلیسی در اسناد این فضا قابل توجه است. به همین منظور علاوه بر بازیابی تک‌زبانه، بازیابی اطلاعات بین-زبانی که در آن پرس‌وجو در زبان مبدا و اسناد در زبان مقصد هستند، مورد توجه قرار گرفت. بازیابی اطلاعات بین‌زبانی به کاربر اجازه می‌دهد که پرس‌وجوی خود را در زبانی غیر از زبان اسناد انتخاب کند. بنابراین، نیاز به منابع ترجمه‌ای خواهد بود که فاصله‌ی بین زبان پرس‌وجو و اسناد را از بین ببرد. از منابع ترجمه‌ای که استفاده می‌شود لغت نامه‌های دوزبانه هستند که با مشکلاتی از قبیل پوشش ندادن همه‌ی کلمات و همچنین واژه‌های جدید روبرو هستند. به همین دلیل، از پیکره‌های دوزبانه[1] برای ساخت ماشین‌های ترجمه آماری[2]، به منظور افزایش کارآیی سیستم‌های بازیابی اطلاعات بین‌زبانی استفاده می‌شود. استفاده از ماشین‌های ترجمه برروی روش‌های اکتشافی سیستم‌های بازیابی اطلاعات، همانند بسامد رخداد کلمه[3] و همچنین بسامد رخداد اسناد[4] اثرگذار خواهد بود.

در سال ۲۰۰۴ Fang و همکاران در مقاله [5] به بررسی تئوری روش‌های اکتشافی، در بازیابی اطلاعات تک‌زبانه پرداختند و محدودیت‌هایی[5] را عنوان کردند که هر سیستم بازیابی اطلاعاتی برای بهبود کارآیی خود باید آن‌ها را برآورده کند. بنابراین، امکان مقایسه‌ی تئوری روش‌های بازیابی اطلاعات تک‌زبانه با توجه به محدودیت‌های تعریف شده فراهم گردید. پس از سال ۲۰۰۴ تمرکز بر روی روش‌های بازیابی اطلاعات تک‌زبانی با رویکردی که عنوان شد ، قرار گرفت و بخش‌های متفاوتی از این سیستم‌ها که در آن‌ها روش‌های اکتشافی در گذشته ارائه شده بود، به صورت تئوری مورد مطالعه قرار گرفت.

همزمان با گسترش وب و مورد توجه قرار گرفتن بازیابی اطلاعات بین‌زبانی، روش هایی برای این سیستم‌ها پیشنهاد و عملیاتی شدند که در آن‌ها روش‌های اکتشافی همانند بازیابی اطلاعات تک‌زبانی نیز عنوان شدند. با توجه به مسائل عنوان شده، این موضوع به ذهن خواهد رسید که بتوان به صورت تئوری نیز روش‌های موجود در بازیابی اطلاعات بین-زبانی و چندزبانی را مورد بررسی و ارزیابی قرار داد و اینکه آیا می‌توان محدودیت‌های عنوان شده در بازیابی اطلاعات تک‌زبانه را، به سیستم‌های بازیابی اطلاعات بین‌زبانی و چندزبانی گسترش داد؟ مطالعه‌ی تئوری روش‌های بازیابی اطلاعات بین‌زبانی، امکان مقایسه‌ی تئوری این روش‌ها را نیز خواهد داد.

با توجه به این که در بازیابی اطلاعات ممکن است پرس‌وجوی کاربر حاوی اطلاعات کافی برای بازیابی اسناد نباشد روش‌هایی با عنوان بازخورد ارتباطی[6] پیشنهاد شدند که در این روش‌ها تلاش بر این است که پرس‌وجوی کاربر را با توجه به اطلاعاتی که به صورت صریح[7]، ضمنی[8] و بازخورد شبه ارتباطی[9] از کاربر و اسناد به دست می‌آورند، گسترش دهند و با استفاده از این کار، کارآیی سیستم‌های بازیابی اطلاعات را افزایش دهند. در روش‌های بازخورد شبه ارتباطی برخلاف دو روش دیگر، نیاز به تلاشی از جانب کاربر نخواهد بود و این روش‌ها با استفاده از کلمات اسناد با رتبه‌ی بالا، پرس‌وجوی کاربر را افزایش می‌دهند. در مقاله [21] روش‌های بازخورد شبه ارتباطی در سیستم‌های بازیابی اطلاعات تک‌زبانه به صورت تئوری مورد مطالعه قرار گرفته است و با استفاده از محدودیت‌هایی که برای این روش‌ها عنوان می-کنند امکان مقایسه‌ی تئوری و همچنین بهبود برخی از این روش‌ها را فراهم می‌آورند. بسامد رخداد کلمه در روش‌های

[1] Bilingual corpora
[2] Statistical translation machines
[3] Term Frequency
[4] Document Frequency
[5] Constraints
[6] Relevance feedback
[7] Explicit
[8] Implicit
[9] Pseudo relevance feedback

بازخورد شبه ارتباطی مورد استفاده قرار خواهد گرفت و همان‌گونه که پیشتر اشاره شد، در بازیابی اطلاعات بین‌زبانی استفاده از ماشین‌های ترجمه بر روی روش‌های اکتشافی در این سیستم‌ها، همانند بسامد رخداد کلمه و بسامد رخداد اسناد اثرگذار خواهد بود. بنابراین مطالعه‌ی تئوری روش‌های بازخورد شبه ارتباطی در سیستم‌های بازیابی اطلاعات بین‌زبانی یکی از بخش‌هایی خواهد بود که در این پژوهش به آن توجه خواهیم کرد.

روشها و فنون اجرایی طرح

با توجه به موارد عنوان شده در قسمت قبل، قصد داریم به صورت تئوری روش‌های بازیابی اطلاعات بین‌زبانی را با محدودیت‌های بیشتری مورد بررسی قرار دهیم و معایب و مزایای هر کدام را بیابیم و سعی در بهبود آن ها خواهیم کرد. رفع نواقص روش‌های بازیابی اطلاعات بین‌زبانی، باعث خواهد شد که شاهد افزایش کارآیی این روش‌ها باشیم. بررسی اینکه آیا می‌توان محدودیت‌های معرفی شده در بازیابی اطلاعات تک‌زبانه را به محدودیت‌هایی برای سیستم‌های بازیابی اطلاعات بین‌زبانی گسترش داد، نیز از دیگر جنبه‌های این پژوهش خواهد بود. علاوه بر آن سعی خواهیم کرد که توابع بسامد رخداد کلمه در بازیابی اطلاعات بین‌زبانی را مورد بررسی قرار داده و مزایا و معایب هر کدام را نشان دهیم. در واقع با توجه به محدودیت‌های تعریف شده، توابع موجود برای بسامد رخداد کلمه را اعتبار سنجی خواهیم کرد .

با توجه به اینکه بازیابی اطلاعات چندزبانی[10] در واقع نوعی از بازیابی اطلاعات بین‌زبانی است، در ادامه‌ی این پژوهش، به بررسی قابلیت گسترش محدودیت‌های بازیابی اطلاعات بین‌زبانی به این سیستم‌ها و تفاوت‌های آن‌ها خواهیم پرداخت.

همان طور که قبلا گفته شد، با توجه به تغییر ماهیتی بازیابی اطلاعات بین‌زبانی با تک‌زبانی قصد داریم ابتدا به صورت تجربی، مقایسه‌ای از نتایج روش‌های بازخورد شبه ارتباطی در بازیابی اطلاعات بین‌زبانی داشته باشیم. با توجه به مقایسه‌ی آماری، می‌توان تفاوت رفتاری روش‌های مختلف بازخورد شبه ارتباطی را در بازیابی اطلاعات بین‌زبانی آشکار کرد. پس از آنکه توانستیم آماری از عملکرد روش‌های بازخورد شبه ارتباطی در بازیابی اطلاعات بین‌زبانی به دست بیاوریم قصد داریم که این روش‌ها را به صورت تئوری مورد بررسی قرار داده و معایب و مزایای هر کدام را مشخص کنیم. شایان ذکر است که این کار در بازیابی اطلاعات تک‌زبانی انجام شده است و قابل تعمیم به بازیابی اطلاعات بین-زبانی خواهد بود. برای این کار نیاز داریم که محدودیت‌هایی را که هر روش بازخورد شبه ارتباطی به صورت معمول باید برآورده کند را مشخص و معرفی کرده و سپس هر کدام از روش‌های بازخورد شبه ارتباطی را با توجه به این محدودیت‌های تعریف شده مورد ارزیابی قرار دهیم و در صورتی که هر کدام از این روش‌ها نواقصی دارند تلاش خواهیم کرد که این نواقص را برطرف کرده و کارآیی آن‌ها را افزایش دهیم.

معیارهای ارزیابی که برای این پژوهش می‌تواند مورد استفاده قرار بگیرد عبارت خواهند بود از: Mean average precision (MAP)، Precision at top k documents (P@k). برای انجام آزمایشات در این پژوهش از داده‌های متنی CLEF شامل اسناد اسپانیایی(۲۰۰۲)، آلمانی(۳-۲۰۰۲)، فرانسوی(۳-۲۰۰۲)، فارسی(۹-۲۰۰۸)، و پرس‌وجوهایی در زبان انگلیسی، استفاده خواهد شد.

[10] Multilingual information retrieval

پیشینه تحقیق (همراه با ذکر منابع اساسی)

هدف از بازیابی اطلاعات بین‌زبانی، امتیازدهی به یک سند در مجموعه اسناد با توجه به یک پرس‌وجو در زبانی غیر از زبان سند است. بنابراین مشکل اصلی موجود در سیستم‌های بازیابی اطلاعات بین‌زبانی، ایجاد یک نگاشت بین کلمات موجود در فضای اسناد و پرس‌وجو است و این نگاشت در واقع یک فرآیند ترجمه را نیاز خواهد داشت. فرآیند ترجمه می‌تواند به سه طریق مورد استفاده قرار گیرد: ۱- ترجمه اسناد به زبان پرس‌وجو ۲- ترجمه پرس‌وجو به زبان اسناد ۳- ترجمه پرس‌وجو و اسناد به یک زبان میانی. در حالت کلی ترجمه کردن پرس‌وجو بیشتر مورد توجه قرار می‌گیرد. اگرچه این روش نیز دارای مشکلاتی از قبیل ابهام در پرس‌وجو است که بر روی ترجمه‌ی پرس‌وجو اثرگذار خواهد بود. پس برای غلبه بر این مشکل از ماشین‌های ترجمه استفاده می شود [۱,۲,۱۶].

Axiomatic analysisای که توسط Fang و همکارانش معرفی شد [۵,۶] بر اساس یک سری محدودیت‌هایی بود که هر سیستم بازیابی منطقی باید آن‌ها را برآورده می‌کرد. محدودیت‌های تعریف شده شامل جنبه‌های مختلف یک سیستم بازیابی اطلاعات بودند: بسامد رخداد کلمه، بسامد رخداد سند، طول سند [۵,۶,۷,۸]، ارتباط ترم‌ها از لحاظ مفهومی [۹,۱۰] و معیارهای ارزیابی [۱۱,۱۲]. در [۱۳] محدودیت‌هایی برای حدپایین بسامد رخداد کلمه معرفی شده است. همچنین در [۲۵] به محدودیت‌هایی برای مجاورت کلمات پرس‌وجو در اسناد پرداخته شده است. در [۲۶] محدودیت‌هایی برای مدل‌های ترجمه معرفی شده است. تمامی این مطالعات بر روی سیستم‌های بازیابی اطلاعات تک‌زبانی بوده است. اگر چه کارهای بسیار زیادی بر روی تحلیل تئوری سیستم‌های بازیابی اطلاعات تک‌زبانی انجام شده است ولی برای تحلیل تئوری سیستم‌های بازیابی اطلاعات بین‌زبانی کارهای بسیار محدودی انجام شده است. در سال ۲۰۱۴ Rahimi و همکارانش به بررسی اولیه دو سیستم بازیابی اطلاعات بین‌زبانی به صورت تئوری پرداخته‌اند. در این مقاله سه محدودیت تعریف شده و از آن‌ها استفاده شده است [۱۳].

بازخورد ارتباطی اولین بار در مدل فضای برداری[11] توسط Rocchio مورد مطالعه قرار گرفت [۱۴]. زمانی که قضاوت‌های ارتباط، در دسترس باشند بازخورد ارتباطی می‌تواند به صورت موثری مورد استفاده قرار بگیرد. به همین علت بازخورد ارتباطی نیاز به تلاش زیادی برای کاربر خواهد داشت که بتواند اسناد با رتبه بالا را امتیازدهی کند. در حالی که در بازخورد شبه ارتباطی فقط نیاز به تعداد اسناد مرتبط خواهیم داشت [۱۷,۱۸]. از روش‌هایی که برای بازخورد شبه ارتباطی، به صورت موثرتری عمل کرده‌اند می‌توان کمینه‌سازی واگرایی[12]، مدل ترکیبی[13]، مدل ارتباطی[14] [۲۳] و مشتقات معرفی شده برای این روش‌ها را نام برد [۲۰,۲۱,۲۲].

در مقاله [۲۱] به صورت تئوری به بررسی روش‌های بازخورد شبه ارتباطی پرداخته‌اند و همچنین محدودیت‌هایی را عنوان کرده‌اند که روش‌های بازخورد شبه ارتباطی باید آن‌ها را برآورده کنند. سپس این مقاله به بررسی تئوری چندین روش بازخورد شبه ارتباطی در بازیابی اطلاعات تک‌زبانی پرداخته است. در مقاله [۲۷] به این موضوع پرداخته شده است که آیا بسامد رخداد اسناد در روش‌های بازخورد شبه ارتباطی تاثیرگذار خواهند بود؟در مقاله [۲۸] نیز به بررسی و تحلیل تاثیر هموارسازی در روش‌های بازخورد شبه ارتباطی پرداخته شده است.

منابع

[1] Ruiz, M., Diekema, A., and Sheridan, P. (2000). "CINDOR Conceptual Interlingua Document Retrieval: TREC-8 Evaluation," in Proceedings of TREC Conference.

[2] Kishida, K., and Kando, N. (2005). "Hybrid approach of query and document translation with pivot language for cross-language information retrieval," in

[11] Vector space model
[12] Divergence minimization
[13] Mixture model
[14] Relevance model

Proceedings of CLEF Conference.

[3] G. Grefenstette, J. Nioche, "Estimation of English and non-English Language Use on the WWW," 6 chemin de Maupertuis, Meylan, France, 2000.

[4] K. Kishida, "Technical issues of Cross-Language Information Retrieval: A Review," *Information Proccessing and Management, pp.433-455, June, 2004*.

[5] H. Fang, T. Tao, and C. Zhai. A formal study of information retrieval heuristics. In SIGIR, pages 49–56. ACM, 2004.

[6] H. Fang and C. Zhai. An exploration of axiomatic approaches to information retrieval. In SIGIR, pages 480–487. ACM, 2005.

[7] S.-H. Na, I.-S. Kang, and J.-H. Lee. Improving term frequency normalization for multi-topical documents and application to language modeling approaches. In ECIR, 2008.

[8] Y. Lv and C. Zhai. Lower-bounding term frequency normalization. In CIKM, pages 7–16. ACM, 2011.

[9] H. Fang and C. Zhai. Semantic term matching in axiomatic approaches to information retrieval. In SIGIR, 2006.

[10] M. Karimzadehgan and C. Zhai. Axiomatic analysis of translation language model for information retrieval. In ECIR, pages 268–280. Springer-Verlag, 2012.

[11] E. Amig´o, J. Gonzalo, and F. Verdejo. A general evaluation measure for document organization tasks. In SIGIR, 2013.

[12] L. Busin and S. Mizzaro. Axiometrics: An axiomatic approach to information retrieval effectiveness metrics. In ICTIR, 2013.

[13] R. Rahimi and I. King, “Axiomatic Analysis of Cross-Language Information Retrieval,” pp. 1875–1878, 2014.

[14] J. Rocchio. Relevance feedback in information retrieval. SMART Retrieval System Experiments in Automatic Document Processing, 1971.

[15] J.Y. Nie "Cross-Language Information Retrieval," *Morgan and Claypool Publisher*, vol. 7*(2)*, pp. 29-55, 2010.

[16] L.A. Ballesteros, "Resolving Ambiguity for Cross-Language Information Retrieval: A Dictionary Approach," in Proceedings of ECAI Conference University, September, 2001.

[17] C. Buckley, G. Salton, J. Allan, and A. Singhal. Automatic query expansion using smart:Trec 3. NIST special publication sp, pages 69–69, 1995.

[18] J. Xu and W. B. Croft. Query expansion using local and global document analysis. In SIGIR ’96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, pages 4–11, New York, NY, USA, 1996. ACM.

[19] V. Lavrenko and W. B. Croft. Relevance based language models. In Proceedings of ACM SIGIR 2001, SIGIR ’01, pages 120–127, New York, NY, USA, 2001 ACM.

[20] Y. Lv and C. Zhai. A comparative study of methods for estimating query language models with pseudo feedback. In Proceedings of ACM CIKM 2009, CIKM ’09, pages 1895–1898, New York, NY, USA, 2009 ACM.

[21] S. Clinchant and E. Gaussier. A theoretical analysis of pseudo-relevance feedback models. In Proceedings of the 2013 Conference on the Theory of Information Retrieval, ICTIR ’13, pages 6:6–6:13, New York, NY, USA, 2013. ACM.

[22] Y. Lv and C. Zhai. Revisiting the divergence minimization feedback model. In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM ’14, pages 1863–1866, New York, NY, USA, 2014. ACM.

[23] C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In Proceedings of ACM CIKM 2001, CIKM ’01, pages 403–410, New York, NY, USA, 2001. ACM.

[24] Yuanhua Lv and ChengXiang Zhai. Lower-bounding term frequency normalization. In CIKM ’11, 2011.

[25] T. Tao and C. Zhai, “An exploration of proximity measures in information retrieval,” inSIGIR 2007, pp. 295–302, 2007.

[26] M. Karimzadehgan and C. Zhai. Axiomatic analysis of translation language model for information retrieval. In ECIR, pages 268–280. Springer-Verlag, 2012.

[27] S. Clinchant and E. Gaussier. Is document frequency important for PRF? In ICTIR, pages 89–100, 2011.

[28] H. Hazimeh and C. Zhai. Axiomatic Analysis of Smoothing Methods in Language Models for Pseudo-Relevance Feedback In ICTIR, 2015.

۵- مصوبه شورای پژوهشی و تحصیلات تکمیلی دانشکده مهندسی برق و کامپیوتر

۵-۱- فرم پیشنهاد و حمایت از پایان‌نامه در تاریخ ..... در شورای پژوهشی و تحصیلات تکمیلی دانشکده /گروه مطرح و نظر شورا به شرح زیر اعلام می‌شود:

☐ **تصویب شد** ☐ **نیاز به اصلاح دارد** ☐ **به تصویب نرسید**

۵-۲- عنوان طرح جامع تحقیقات استاد راهنما: سیستم‌های اطلاعاتی و محیط‌های هوشمند

امضاء استاد راهنما

۵-۳- آیا پایان نامه پیشنهادی مرتبط با طرح جامع تحقیقات استاد راهنما/مشاور/گروه آموزشی/ دانشکده می باشد:

☐ **بلی** ☐ **خیر**

امضاء رئیس / معاون پژوهشی و تحصیلات تکمیلی دانشکده مهندسی برق و کامپیوتر

شماره:
تاریخ:

معاون محترم آموزشی و تحصیلات تکمیلی پردیس دانشکده های فنی
با سلام و احترام،
فرم پیشنهاد و حمایت از پایان‌نامه کارشناسی ارشد آقای علی منتظرالقائم **با عنوان** تحلیل نظری بازیابی اطلاعات بین زبانی **به راهنمایی** خانم دکتر آزاده شاکری **در شورای پژوهشی و تحصیلات تکمیلی دانشکده مهندسی برق و کامپیوتر مورخ ..... به تصویب رسید.**
خواهشمند است دستور فرمایید اقدامات مقتضی انجام شود.

امضاء رئیس / معاون پژوهشی و تحصیلات تکمیلی دانشکده مهندسی برق و کامپیوتر

شماره:

تاریخ:

معاون محترم پژوهشی پردیس دانشکده های فنی

با سلام و احترام ,

به پیوست فرم پیشنهاد و حمایت از پایان‌نامه تحصیلات تکمیلی با مشخصات مذکور که به تصویب شورای پژوهشی وتحصیلات تکمیلی دانشکده مهندسی برق و کامپیوتر رسیده است، جهت دستور اقدام مقتضی تقدیم می‌شود.

امضاء معاون آموزشی و تحصیلات تکمیلی پردیس دانشکده های فنّی

رونوشت: معاون محترم پژوهشی و تحصیلات تکمیلی دانشکده مهندسی برق و کامپیوتر: جهت اطلاع و پیگیری