مقدمه‌ای بر بازیابی اطلاعات

تألیف:

کریستوفر دی. منینگ

پرباکر رَگِوِن

هنریک شوتزه

ترجمه:

دکتر هدیه ساجدی (عضو هیأت علمی دانشگاه تهران)

مهندس زهرا سادات تقوی

فرناز سادات تقوی

نیاز دانش

مقدمه مترجمین

بازیابی اطلاعات به فن‌آوری و دانش پیچیده جستجو و استخراج اطلاعات، داده‌ها و فراداده‌ها در انواع گوناگون منابع اطلاعاتی اطلاق می‌شود. با افزایش روزافزون حجم اطلاعات ذخیره‌شده در منابع مختلف، فرآیند بازیابی و استخراج اطلاعات اهمیت ویژه‌ای پیدا کرده است. برخلاف پایگاه داده‌ها، اطلاعات ذخیره شده در منابع اطلاعاتی بزرگ مانند وب و زیرمجموعه‌های آن مانند شبکه‌های اجتماعی، از ساختار مشخصی پیروی نمی‌کنند و عموماً دارای معانی تعریف شده و مشخصی نیستند. هدف بازیابی اطلاعات در چنین شرایطی، کمک به کاربر برای یافتن اطلاعات مورد نظر در انبوهی از اطلاعات بدون ساختار است.

کتاب مقدمه‌ای بر بازیابی اطلاعات اولین کتاب با درک منسجمی از بازیابی اطلاعات سنتی و وب است که شامل جستجوی وب و حوزه‌های مرتبط با دسته‌بندی و خوشه‌بندی متن است. یکی از دلایل ترجمه این کتاب، انسجام خوب آن در این حیطه بود که توسط اساتید متخصص دانشگاهی نوشته شده است و به خوبی می‌تواند نیاز خوانندگان را مرتفع کند. از طرفی تلاش ما بر این بود که کتاب درسی منسجمی به زبان فارسی ارائه شود که به تحولات نسبتاً زیاد حیطه بازیابی اطلاعات پرداخته باشد.

در بخش پیشگفتار، مؤلفان کتاب، ویژگی‌های کتاب را به خوبی مطرح کرده‌اند و نیازی به تکرار آنها نیست. شایان توجه است که این کتاب در مجموع، با قلمی روان طیف قابل قبولی از مطالب را در زمینه بازیابی اطلاعات مطرح کرده است. به طوریکه طبق مصوبات شورای عالی برنامه‌ریزی وزارت علوم، تحقیقات و فن‌آوری تقریباً کلیه سرفصل‌های درس کارشناسی و کارشناسی ارشد بازیابی اطلاعات پیشرفته را پوشش می دهد.

در ترجمه این کتاب بر امانت‌داری، صحت و روانی ترجمه تاکید شده است. از ابداع واژه‌های جدید و مترادفِ واژه‌های موجود پرهیز و سعی شده از منابع موجود استفاده شود. مجموعه واژگان (فارسی به انگلیسی و انگلیسی به فارسی) در انتهای کتاب آورده شده است.

در اینجا لازم می‌دانیم تشکر ویژه‌ای از اعضای خانواده داشته باشیم. بدیهی است انجام اینکار بدون حمایت و همراهی آنها عملی نبود.

به یقین ترجمه حاضر خالی از اشکال و خطا نیست. بنابراین از تمامی عزیزانی که این کتاب را مطالعه می‌کنند، تقاضا داریم که ما را از نظرات و انتقادات خود از طریق ناشر مطلع نمایند. در انتها امیدواریم که مطالب این کتاب بتواند پاسخگوی نیاز اساتید، دانشجویان و افراد علاقه‌مند به حیطه بازیابی اطلاعات بوده و موجبات تعالی دانش آنها را فراهم آورد.

مترجمین

زمستان ۱۳۹۳

پیشگفتار

مطالعات دهه ۱۹۹۰ نشان دادند که اکثر افراد ترجیح می‌دادند که اطلاعات را از یکدیگر دریافت کنند تا اینکه از سیستم‌های بازیابی اطلاعات بگیرند. البته، در آن دوره زمانی، اکثر افراد از دفاتر آژانس‌های مسافرتی نیز برای رزرو سفر خود استفاده می‌کردند. هر چند، در طول دهه گذشته، بهینه‌سازی‌ها در راستای اثربخشی بازیابی اطلاعات، موتورهای جستجوی وب را ارتقاء داده، طوری که اکثر افراد بیشتر اوقات از آن راضی بوده و جستجوی وب به یک استاندارد تبدیل شده و اغلب مرجع یافتن اطلاعات گشته است. برای مثال، Pew Internet Survey در سال ۲۰۰۴ (Fallows (2004)) بیان کرد که «۹۲٪ کاربران اینترنت بیان می‌کنند که اینترنت جای خوبی برای گرفتن اطلاعات روزانه است». حوزه بازیابی اطلاعات برای بسیاری از افراد جالب توجه است و از اطلاعات سیستم دانشگاهی تا مبانی اساسی دستیابی به اطلاعات را تحت تأثیر قرار داده است. این کتاب پشتیبانی علمی در این زمینه را در سطحی که برای دانشجویان و مقاطع بالاتر قابل استفاده باشد، فراهم می‌آورد.

بازیابی اطلاعات همزمان با وب آغاز نشد: در پاسخ به چالش‌های متعدد در فراهم آوردن دسترسی به اطلاعات، حوزه بازیابی اطلاعات برای ارائه روش‌های جستجو برای صورت‌های مختلف محتوا مطرح شد. این حوزه با انتشارات علمی و رکوردهای کتابخانه‌ای آغاز گشت، اما خیلی زود به صورت‌های دیگر محتوا، به ویژه آن صورت‌هایی که مورد استفاده متخصصان است مانند، روزنامه-نگاران، موکّلان، و پزشکان است بسط پیدا کرد. اکثر تحقیقات علمی در بازیابی اطلاعات روی این محتواها رخ داده است و اکثر کاربردهای ادامه‌دار از بازیابی اطلاعات، به فراهم آوردن دسترسی به اطلاعات بدون ساختار در شرکت‌ها و دامنه‌های دولتی مختلف سروکار داشته و این کاربردها و مباحث بخش اساسی این کتاب را تشکیل می‌دهد.

با این‌حال، در سال‌های اخیر، محرک اصلی بازیابی اطلاعات، شبکه جهانی وب بوده است که در مقیاس ده‌ها میلیون تولیدکننده محتوا عمل کرده است. اگر اطلاعات بازیابی و تفسیر و تحلیل نمی‌شد طوریکه کاربر بتواند به سرعت اطلاعاتی مرتبط به نیازش را بیابد این انفجار اطلاعات منتشر شده مطرح نمی‌شد. در اواخر دهه ۱۹۹۰، افراد زیادی دریافتند که ادامه شاخص‌گذاری کل وب، به دلیل رشد نمایی وب از نظر اندازه، به سرعت غیرممکن خواهد شد. اما بخش اعظم نوآوری‌های علمی و مهندسی، کاهش سریع قیمت سخت افزار کامپیوتر و افزایش تجارت تحت جستجوی وب، همگی به

قدرت امروزی اکثر موتورهای جستجو افزودند. این موتورها قادر هستند، نتایج با کیفیت بالا را در زمان پاسخ کمتر از یک ثانیه برای هزاران میلیون جستجو در روز، روی میلیاردها صفحه وب فراهم کنند.

سازماندهی کتاب و توسعه دوره

این کتاب نتیجه یک سری دوره‌هایی است که در دانشگاه استنفورد و دانشگاه اشتوتگارت در بازه زمانی یک نیم ترم، یک ترم کامل و دو نیم ترم تدریس کرده‌ایم. این دوره‌ها بیشتر برای دانشجویان در اوایل دوره کارشناسی ارشد علوم کامپیوتر برگزار شده است. اما دانشجویان وکالت، پزشکی، آمار، زبان‌شناسی و گرایشات مختلف مهندسی را نیز شامل شده است. از اینرو، اصل کلیدی طراحی این کتاب، پوشش مطالبی بود که اعتقاد داریم در دوره یک ترم تحصیلی در بازیابی اطلاعات مهم است. اصل دیگر ایجاد هر فصل براساس مطالبی است که می‌توان آن را در ارائه درس ۷۵ تا ۹۰ دقیقه‌ای پوشش داد.

هشت فصل اول کتاب، به مبانی بازیابی اطلاعات و به ویژه قلب موتورهای جستجو تخصیص داده شده، که ما این مطالب را اساس هر دوره‌ای در بازیابی اطلاعات می‌دانیم. فصل ۱، شاخص‌های وارونه را معرفی می‌کند و نشان می‌دهد که چگونه پرس و جو های بولی ساده می‌تواند با استفاده از چنین شاخص‌هایی پردازش شود. فصل ۲، این مقدمه را با ارائه جزئیات روشی که در آن اسناد قبل از شاخص‌گذاری پیش پردازش شده و با بحث راجع به بکارگیری شاخص وارونه به شیوه‌های مختلف برای افزایش سرعت و عملیاتی بودن، بسط می‌دهد. فصل ۳، ساختارهای جستجو برای لغت‌نامه‌ها و چگونگی پردازش پرس و جوهایی که خطاهای املایی داشته و دیگر تطبیق‌های غیردقیق به مجموعه واژگان در مجموعه اسنادی که در حال جستجو است را مورد بحث قرار می‌دهد. فصل ۴، شماری از الگوریتم‌ها را برای ساخت شاخص وارونه از مجموعه متن با توجه ویژه به الگوریتم‌های توزیع شده و مقیاس‌پذیر توصیف می‌کند، که می‌توانند در مجموعه‌های بسیار بزرگ به کار گرفته شوند. فصل ۵، روش‌های فشرده‌سازی لغت‌نامه و شاخص‌های وارونه را پوشش می‌دهد. این روش‌ها، برای دستیابی به زمان‌های پاسخ کمتر از یک ثانیه برای پرس و جوهای کاربر در موتورهای جستجوی بزرگ، ضروری هستند. شاخص‌ها و پرس و جوهایی که در فصول ۱ تا ۵ مورد توجه قرار گرفتند، تنها با بازیابی بولی سروکار دارند که در آن یک سند یا با یک پرس و جو تطبیق دارد یا تطبیق ندارد. تمایل به اندازه‌گیری میزانی که یک سند با یک پرس و جو تطبیق دارد، یا نمره یک سند برای یک پرس و جو، موجب طرح وزن‌دهی عبارت و محاسبات نمره‌ها در فصول ۶ و ۷ می‌شود و به ایده‌ی لیستی از اسناد منتهی می‌شود که برای پرس و جو رتبه بندی شده‌اند. فصل ۸، بر ارزیابی سیستم بازیابی اطلاعات مبتنی بر ربط سندی که آن را بازیابی می‌کند، تمرکز کرده و به ما اجازه می‌دهد تا عملکردهای نسبی سیستم‌های مختلف را براساس مجموعه‌های سند و پرس و جو که برای محک و ارزیابی سیستم‌های بازیابی اطلاعات ارائه شده‌اند مقایسه کنیم.

فصول ۹ تا ۲۱ بر مبنای هشت فصل اول، انواع موضوعات پیشرفته‌تر را پوشش می‌دهند . فصل ۹ روش‌هایی را بررسی می‌کند که توسط آن بازیابی می‌تواند از طریق روش‌هایی مانند، بازخورد ربط و توسعه پرس‌وجو، که برای افزایش درستنمایی بازیابی اسناد مرتبط تلاش می‌کنند، بهبود یابد. فصل ۱۰، بازیابی اطلاعات از اسنادی را در نظر می‌گیرد که با زبان‌های نشانه‌گذاری مانند XML و HTML ساختار یافته‌اند. بازیابی ساخت‌یافته را به وسیله کاهش آن به روش‌های نمره‌گذاری فضای بردار که در فصل ۶ مطرح شده ارائه می‌دهیم. فصول ۱۱ و ۱۲، به نظریه احتمالاتی برای محاسبه نمره‌های سند در پرس و جو استناد می‌کند. فصل ۱۱، بازیابی اطلاعات احتمالاتی سنتی را توسعه می‌دهد که چارچوبی را برای محاسبه احتمال ربط یک سند با معلوم بودن مجموعه عبارات پرس و جو فراهم می‌آورد. از اینرو، این احتمال می‌تواند به عنوان نمره‌ای در رتبه بندی استفاده شود. فصل ۱۲، راه حل دیگری را شرح می‌دهد که در آن برای هر سندی در مجموعه، یک مدل زبانی می‌سازیم که از آن می‌توان احتمالی که مدل زبانی یک پرس و جوی مشخص را ایجاد کند، تخمین بزنیم. این احتمال کمّیت دیگری است که با آن می‌توان اسناد را رتبه بندی کرد.

فصول ۱۳ تا ۱۷ صورت‌های مختلف یادگیری ماشین و روش‌های عددی در بازیابی اطلاعات را بیان می‌کنند. فصول ۱۳ تا ۱۵ مسئله دسته بندی اسناد را با دانستن مجموعه اسناد و دسته‌هایی که به آن تعلق دارند، بیان می‌کند. فصل ۱۳، دسته بندی آماری را به عنوان یکی از تکنولوژی‌های کلیدی برای موتورجستجو مطرح کرده و Naïve Bayes را که روش ساده و کارآمد دسته بندی متن است، معرفی می‌کند و روش استانداردی را برای ارزیابی دسته بندی متن طرح ریزی می‌کند. فصل ۱۴، مدل فضای بردار را از فصل ۶ اتخاذ کرده و دو روش دسته بندی Rocchio و k–نزدیکترین همسایه (kNN) را معرفی می‌کند. همچنین، مصالحه بایاس–واریانس را به عنوان ویژگی مهمی از مسائل یادگیری مطرح می‌کند که ضابطه‌ای را برای انتخاب یک روش مناسب برای مسئله دسته بندی متن فراهم می‌آورد. فصل ۱۵، ماشین‌های بردار پشتیبان را معرفی می‌کند که بسیاری از محققان آن را در حال حاضر به عنوان کارآمدترین روش دسته بندی معرفی می‌کنند. همچنین، روابطی را بین مسئله دسته‌بندی و موضوعات به ظاهر مجزا، مانند استنتاج توابع نمره‌گذاری از مجموعه مثال‌های آموزشی، ارائه می‌دهیم.

فصول ۱۶ تا ۱۸ مسئله ایجاد خوشه‌ها را از اسناد مرتبط در یک مجموعه در نظر می‌گیرند. در فصل ۱۶، ابتدا مروری بر شماری از کاربردهای مهم خوشه‌بندی در بازیابی اطلاعات خواهیم داشت. سپس دو الگوریتم خوشه بندی را معرفی می‌کنیم: الگوریتم K-means که به طور گسترده و مؤثر برای خوشه‌بندی اسناد به کار می‌رود؛ و الگوریتم بیشینه‌سازی امیدریاضی[1] که از نظر ریاضی پر هزینه‌تر بوده اما منعطف‌تر می‌باشد. فصل ۱۷، خوشه‌بندی سلسله مراتبی ساخت یافته را در بسیاری از کاربردهای بازیابی اطلاعات مطرح کرده و شماری از الگوریتم‌های خوشه‌بندی که سلسله مراتب خوشه‌ها را تولید می‌کنند، معرفی می‌کند. این فصل، همچنین مسئله دشوار محاسبه خودکار برچسب‌ها

[1] Expectation Maximization

را برای خوشه‌ها مطرح می‌کند. فصل ۱۸، به طرح روش‌های جبرخطی که در خوشه‌بندی می‌توانند مورد استفاده قرار گیرند پرداخته و همچنین کاربردهای روش‌های جبری در بازیابی اطلاعات را مطرح می‌کند که در روش شاخص‌گذاری معنایی نهان بکار گرفته شده‌اند.

فصول ۱۹ تا ۲۱، مسئله جستجوی وب را در نظر می‌گیرند. در فصل ۱۹، بطور خلاصه به چالش‌های پایه در جستجوی وب به همراه مجموعه‌ای از روش‌ها در بازیابی اطلاعات وب پرداخته می‌شود. سپس، فصل ۲۰، معماری و الزامات پیمایشگر وب را توصیف می‌کند. در نهایت فصل ۲۱، قدرت تحلیل پیوند[1] را در جستجوی وب در نظر می‌گیرد که در این روال، چندین روش را از جبر خطی گرفته تا نظریه احتمالات پیشرفته به کار می‌گیرد.

این کتاب، تمامی عناوین مرتبط با بازیابی اطلاعات را بطور جامع پوشش نمی‌دهد. ما شماری از موضوعات را کنار گذاشتیم زیرا خارج از محدوده مطالبی است که تمایل به پوشش آنها در مقدمه‌ای بر بازیابی اطلاعات داشتیم. با این حال، برای افرادی که علاقه‌مند به این موضوعات هستند، چندین کتاب معرفی می‌کنیم.

بازیابی اطلاعات بین زبانی؛ Grossman and Frieder (2004)، فصل ۴؛ و Oard and Dorr (1996).

بازیابی اطلاعات تصویر و چندرسانه‌ای؛ Grossman and Frieder (2004)، فصل ۴؛ Baeza-Yates and Ribeiro-Neto (1999)، فصل ۶؛ Baeza-Yates and Ribeiro-Neto (1999)، فصل ۱۱؛ Baeza-Yates and Ribeiro-Neto (1999)، فصل ۱۲؛ del Bimbo (1999)؛ Lew (2001)؛ و Smeulders et al. (2000).

بازیابی صوت؛ Coden et al. (2002).

بازیابی موسیقی؛ Downie (2006) و http://www.ismir.net/ .

واسط‌های کاربری برای بازیابی اطلاعات؛ Baeza-Yates and Ribeiro-Neto (1999)، فصل ۱۰.

بازیابی اطلاعات موازی و نظیر به نظیر؛ Grossman and Frieder (2004)، فصل ۷؛ Baeza-Yates and Ribeiro-Neto (1999)، فصل ۹؛ و Aberer (2001).

کتابخانه‌های دیجیتال؛ Baeza-Yates and Ribeiro-Neto (1999)، فصل ۱۵؛ و Lesk (2004).

چشم‌انداز علوم اطلاعات؛ Korfhage (1997)، Meadowet al. (1999)، و Ingwersen and Järvelin (2005).

روش‌های مبتنی بر منطق، برای بازیابی اطلاعات؛ van Rijsbergen (1989).

روش‌های پردازش زبان طبیعی؛ Manning and Schütze (1999)، Jurafsky and Martin (2008) و Lewis and Jones (1996).

پیش نیازها

دوره‌های مقدماتی در ساختمان داده‌ها و الگوریتم‌ها، جبر خطی و نظریه احتمالات به عنوان پیش نیاز برای تمامی ۲۱ فصل کفایت می‌کند.

[1] Link Analysis

فصول ۱ تا ۵، دوره مبانی الگوریتم و ساختمان داده‌ها را به عنوان پیش نیاز در نظر می‌گیرد. به علاوه، فصول ۶ و ۷ به دانش جبرخطی پایه شامل بردارها و ضرب نقطه‌ای نیاز دارند. هیچ پیش نیاز دیگری تا فصل ۱۱ لازم نیست. در فصل۱۱ دوره مبانی در نظریه احتمالات مورد نیاز می‌باشد؛ بخش ۱۱–۱ مرور کوتاهی دارد بر مفاهیم مورد نیاز فصول ۱۱ تا ۱۳. فصل ۱۵ فرض می‌کند که خواننده با مفهوم بهینه‌سازی غیرخطی آشنا است، اگر چه این فصل ممکن است بدون دانش دقیق در مورد الگوریتم‌های بهینه‌سازی غیرخطی نیز مطالعه شود. فصل ۱۸ نیازمند دوره مقدماتی در جبرخطی شامل آشنایی با مفاهیم رتبه ماتریس و بردار مشخصه است که مرور مختصری بر آن در بخش ۱۸–۱ ارائه شده است. دانش مرتبط با مقادیر ویژه و بردارهای ویژه نیز در فصل ۲۱ مورد نیاز است. سطح دشواری تمرینات با آسان (★)، متوسط (★★)، و یا دشوار (★★★) مشخص می‌گردد.

تقدیر و تشکر

از انتشارات دانشگاه کمبریج تشکر می‌کنیم که به ما اجازه دادند پیش‌نویسی از کتاب را به صورت آنلاین ایجاد کنیم و بازخوردی از کتاب در حین نگارش آن به دست آوریم. همچنین از Lauren Cowles تشکر می‌کنیم. او ویراستار برجسته‌ای است و چندین دوره نظراتی در مورد هر فصل ازنظر شیوه نگارش، سازماندهی و پوشش مطالب، ارائه کرد. وی در رسیدن ما به اهداف نگارش این کتاب نقش ویراستاری بسزایی داشته است.

ما مدیون افراد بسیاری هستیم که نظرات، پیشنهادات و اصلاحاتی بر مبنای نسخه پیش‌نویس این کتاب ارائه کرده‌اند.

افراد بسیاری بازخورد دقیق در مورد هر یک از فصول یا بنا به درخواست و یا براساس خواست خود ایجاد کرده‌اند و برای این امر قدردان آنها هستیم.

و در آخر از داوران که بر حسب کیفیت و کمّیت نظراتی را فراهم آورده‌اند به خاطر تأثیر قابل توجه آنها بر محتوا و ساختار کتاب تشکر می‌کنیم. ما قدردانی خود را از Pavel Berkhin، Stefan Büttcher، Jamie Callan، Byron Dom، Torsten Suel و Andrew Trotman اعلام می‌داریم.

بخش‌های پیش‌نویس اولیه فصول ۱۳، ۱۴، و ۱۵، براساس اسلایدهایی است که توسط Ray Mooney فراهم شده‌اند. با وجود اینکه مطالب اصلاحات زیادی داشته است، اما از مشارکت Ray در این سه فصل و به ویژه توصیف پیچیدگی زمانی در تمامی الگوریتم‌های دسته‌بندی متن قدردانی می‌کنیم.

مؤلفان از دانشگاه استنفورد و اشتوتگارت برای فراهم آوردن محیط دانشگاهی برای به بحث گذاشتن ایده‌ها و فرصت تدریس دوره‌هایی که از آن این کتاب نتیجه شده و در آن محتوایش اصلاح شده است، تشکر می‌کنند. کریستوفر دی. منینگ از خانواده‌اش برای ساعات بسیاری که وی صرفِ کار بر روی این کتاب کرده است، تشکر کرده و امیدوار است که سال آینده وقت آزاد بیشتری را آخر هفته‌ها با آنها سپری کند. پرباکر رگون از خانواده‌اش برای پشتیبانی صبورانه‌شان در طول نگارش این کتاب قدردانی کرده و به Yahoo! برای فراهم آوردن محیط پرثمری که در آن این کتاب کار شده

است، مدیون است. هنریک شوتزه از والدین، خانواده و دوستانش برای حمایت‌شان در طول نگارش این کتاب تشکر می‌کند.

اطلاعات تماس و وب

این کتاب، وب سایتی با آدرس http://informationretrieval.org دارد. برروی این وب سایت مجموعه‌ای از اسلایدها که برای هر فصل ایجاد شده قرار داده شده است و می‌تواند برای تدریس درس بازیابی اطلاعات بکار رود. ما از بازخوردها، اصلاحات و پیشنهادات استقبال می‌کنیم که می‌تواند از طریق informationretrieval (at) yahoogroups (dot) com. به تمامی مؤلفان فرستاده شود.

فهرست مطالب