

پردازش زبان طبیعی و بازیابی اطلاعات

ترجمه

دکتر جعفر مهراذ - مریم ناصری



وزارت فرهنگ و ارشاد اسلامی
مرکز اسناد و کتابخانه ملی جمهوری اسلامی ایران



سرشناسه	: مهرداد، جعفر، ۱۳۲۵ - گردآورنده، مترجم.
عنوان و نام پدیدآور	: پردازش زبان طبیعی و بازیابی اطلاعات / [گردآوری] و ترجمه جعفر مهرداد، مریم ناصری.
مشخصات نشر	: تهران: چاپار: شیراز: مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری، ۱۳۸۷.
مشخصات ظاهری	: ز: ۱۱۰ ص: ۱ جدول.
شابک	: ISBN: 978-600-5294-12-5
وضعیت فهرست‌نویسی: فبا.	
یادداشت	: کتاب حاضر گزیده‌ای از مطالعاتی است که در مراکز علمی جهان در حیطه زبان فارسی، اسپانیولی و انگلیسی صورت گرفته است.
یادداشت	: واژه‌نامه.
یادداشت	: نمایه.
یادداشت	: کتاب‌نامه.
موضوع	: ذخیره و بازیابی اطلاعات
موضوع	: پردازش زبان طبیعی
موضوع	: نمایه‌سازی
شناسه افزوده	: ناصری، مریم، ۱۳۶۳ - گردآورنده، مترجم
رده‌بندی کنگره	: ۱۳۸۷ پ ۴۹۹۹/م ۹۹۹
رده‌بندی دیویی	: ۰۲۵/۰۴
شماره کتابشناسی ملی	: ۱۵۳۹۶۲۱



سازمان اسناد و کتابخانه ملی
جمهوری اسلامی ایران



تهران: ۸۸۸۹۹۶۸۰ - ۰۹۱۳۱۹۶۰۲۱۴

پردازش زبان طبیعی و بازیابی اطلاعات

ترجمه دکتر جعفر مهرداد و مریم ناصری

طراح جلد و صفحه‌آرا: کریم فلاح

چاپ اول: ۱۳۸۷

شمارگان: ۲۰۰۰

ناظر فنی چاپ: مریم تفضلی

شابک: ۹۷۸-۶۰۰-۵۲۹۴-۱۲-۵ ۹۷۸-۶۰۰-۵۲۹۴-۱۲-۵ ISBN: 978-600-5294-12-5

حق هرگونه چاپ و انتشار محفوظ است.

قیمت: ۲۰۰۰ تومان

فهرست

صفحه	عنوان
۱ - ۱۶	فصل اول
	پردازش زبان طبیعی در بازیابی اطلاعات
۱	مقدمه
۲	روش های ارزیابی بازیابی
۳	پردازش زبان طبیعی در بازیابی اطلاعات
۱۰	تاثیر طول پرس وجو
۱۰	پردازش زبان طبیعی و TREC
۱۲	نتایج
۱۷ - ۳۰	فصل دوم
	پردازش زبان طبیعی در بازیابی اطلاعات متنی
۱۷	مقدمه
۱۸	مسایل پردازش زبان طبیعی: اختلاف زبانی و ابهام
۱۹	پردازش زبان طبیعی در بازیابی اطلاعات متنی
۲۶	عناوین مربوط در تحقیقات
۲۷	نتایج
۳۱ - ۴۰	فصل سوم
	نقش پردازش زبان طبیعی در نظام بازیابی اطلاعات هوشمند
۳۱	مقدمه
۳۲	نظام بازیابی اطلاعات قراردادی و نظام بازیابی هوشمند
۳۵	زبان طبیعی در بازیابی اطلاعات هوشمند
۳۵	نقش پردازش زبان طبیعی در بازیابی اطلاعات هوشمند
۳۸	نتیجه

بازیابی اطلاعات با دقت بیشتر با استفاده از فنون پردازش زبان طبیعی

۴۱	مقدمه
۴۲	پردازش زبان طبیعی
۴۳	مجموعه‌های داده‌ها
۴۴	فضای برداری
۴۶	همجواری
۴۸	الگوریتم پشتیبانی
۴۸	شناسگر عبارت اسمی
۴۹	ترکیب کردن مدل‌ها
۵۰	نتایج

الگوریتم ریشه‌سازی برای زبان فارسی

۵۳	مقدمه
۵۴	زبان فارسی
۵۶	الگوریتم
۵۸	پیاده سازی (اجراء)
۶۱	ارزیابی
۶۲	نتیجه و تحقیقات بیشتر

جستجوی فارسی و فناوری‌های نمایش

۶۵	مقدمه
۶۶	سیستم درون‌داد و برون‌داد فارسی

۶۸	اپلت صفحه کلید
۶۹	معماری سیستم
۷۱	ریشه ساز و فهرست واژه های غیر مجاز
۷۴	مجموعه اسناد
۷۴	موتور جستجو
۷۶	نتیجه و مطالعات آتی
۷۹ - ۸۹	فصل هفتم
	ریشه سازی و اِن-گرام ها در اسپانیولی و تاثیر آنها در بازیابی اطلاعات
۷۹	مقدمه
۸۱	اِن-گرام ها
۸۱	بازیابی اطلاعات و ریشه سازی در اسپانیولی
۸۲	مجموعه مدارک
۸۲	مدل بازیابی
۸۳	ریشه سازی
۸۴	اِن-گرامها در اسپانیولی
۸۵	نتایج
۹۱ - ۹۹	فصل هشتم
	پردازش زبان طبیعی عربی برای بازیابی اطلاعات
۹۱	مقدمه
۹۳	مهمترین مشخصه های زبان عربی
۹۴	فنون ریشه سازی
۹۵	آزمون های آماری
۹۶	موتورهای جستجوی وبی عربی و معیارهای ارزیابی
۹۸	نتیجه

واژه نامه انگلیسی - فارسی

نمایه

۱۰۱ - ۱۰۷

۱۰۹ - ۱۱۰

از بعد از جنگ جهانی دوم، بازیابی اطلاعات به برنامه‌ای مهم در دانشکده‌های کتابداری و اطلاع رسانی تبدیل شد و با شروع استفاده از رایانه‌های شخصی در کتابخانه‌ها و مرکز اطلاع رسانی و تولید برنامه‌های نرم افزاری، توجه به بازیابی اطلاعات در رأس برنامه‌ها و خدمات بنگاه‌های تجاری کامپیوتری قرار گرفت. تولید پایگاه‌های اطلاعاتی تمام متن از دهه‌های ۱۹۷۰ و ۱۹۸۰ میلادی به بعد، به موضوع مورد علاقه پژوهشگرانی مبدل شد که در مبحث بازیابی خودکار متن یا بازیابی مدارک در حوزه پردازش زبان طبیعی فعالیت می‌کنند.

هدف این اثر نشان دادن خواص اساسی بازیابی مدرک، ارائه خلاصه‌ای از تجارب پژوهشگران، نمایاندن مطالعات جدید در حوزه ریشه‌سازی و تعیین مقوله‌های دستوری و تأثیر آن در علایق تحقیقاتی در بازیابی متن، و بررسی راهبردهای ویژه در تحقیقات پردازش زبان طبیعی می‌باشد.

به خاطر توجه به علایق خوانندگان این اثر از اصطلاح قدیمی‌تر به عنوان بازیابی مدرک و از اصطلاح جدیدتر به عنوان بازیابی متن به طور مترادف استفاده شده است. هدف هر دو بازیابی متون است: از یک پاراگراف گرفته تا متن یک کتاب.

در فصولی که برای مطالعه خوانندگان این کتاب انتخاب شده است، توجه ما معطوف به بازیابی اطلاعات، و به طور اخص معطوف به بازیابی مدرک به عنوان فعالیت پردازش زبان طبیعی می‌باشد.

نکته‌ای که حائز اهمیت می‌باشد این است که در قلمرو کلی بازیابی اطلاعات بازیابی مدرک فعالیتی است خاص و مهم که دانشجویان رشته علوم کتابداری و اطلاع رسانی نباید آن را با بازیابی داده‌ها یا دانش اشتباه کنند.

بازیابی اطلاعات به نمایه‌سازی بستگی دارد. نمایه‌سازی به نوبه خود مستلزم یک زبان نمایه‌سازی است با مجموعه‌ای از اصطلاحات و روش‌ها برای ساخت درخواست توصیفگرهای مدرک. نمایه‌سازی اساس بازیابی مدارکی است که به نیاز کاربر ربط دارد. نمایه‌سازی باید به وسیله ابزارهای جستجو که شرایط را برای تطبیق بین درخواست و توصیفگرهای مدرک مشخص می‌سازند، پشتیبانی شود.

اساس نمایه‌سازی افزایش دقت (یعنی، نسبت مدرک بازیابی شده که مربوط می‌باشند) و

بازیافت (یعنی، نسبت مدارک مربوط که بازیابی می شوند) می باشد.

دو نوع مسئله وجود دارد که نمایه سازی باید با توجه به این مسایل به این هدف ها نایل شود:

نخست، مسایلی وجود دارد که به وسیله بافت بیرونی، که جستجو در آن انجام می پذیرد، مطرح می شود. در این حالت، نوعاً تعدادی مدارک مربوط و تعداد زیادی مدارک نامربوط وجود دارد.

دوم، مسایلی وجود دارد که به وسیله محدودیت های درونی خود سیستم اعمال می شود:

- تغییرپذیری در روش هایی که سیستم بازیابی باید بر آنها غلبه کند. این مورد تا حدودی به مسئله زبان مربوط می شود.

- تصریح اندک خصوصیات درخواست که آن را تا حدودی با ابهام مواجه می سازد، یا این که درخواست ممکن است کامل نباشد (تفاوتی که بین این دو وجود دارد آن است که در مورد اول کاربر ممکن است ابهام را تشخیص ندهد و در مورد دوم از ارائه اطلاعات کافی کوتاهی کند).

- توصیفگرها که ممکن است غیر مستقیم باشند چون متون کامل مدارک به طور فزاینده به صورت روی خط در دسترس می باشند، موضوع توصیفگرها در مقایسه با سأل های گذشته بیشتر تحت کنترل روش های نمایه سازی است. با وجود این، همیشه نمی توان توصیفگرها را به طور کامل تهیه دید. نویسنده یک مدرک همیشه گفته های فراوانی دارد که در مدرک به طور کامل ارائه نمی دهد.

از این رو، بازیابی مدرک تکالیف ناسازگاری را در توصیف مدرک تحمیل می کند. نتیجه آن است که اختلافات در نمایه سازی باعث تفاوت هایی می شود که در بازیافت و دقت بازیافت مشاهده می شود. فایق آمدن بر این موضوع و افزایش بازیافت هدف اساسی در ساخت زبان نمایه سازی است.

امکانات و گزینه های زیادی برای زبان های نمایه سازی وجود دارد. اصطلاحات ممکن است اصطلاحاتی باشند که در متن در دست نمایه سازی وجود دارد (زبان طبیعی)، یا ممکن است محدود به یک زبان مصنوعی یا کنترل شده باشد که در سیستم بازیابی مورد استفاده قرار می گیرد و طراحی آن بسیاری از ملاحظات را که در بحث با بازنمون معنی در پردازش زبان

طبیعی ضروری است ایجاب می نماید.

تجربه ما نشان می دهد که علیرغم اهمیت مسایل زبانشناسی در بازیابی اطلاعات متخصصان رشته علوم کتابداری و اطلاع رسانی در دانشگاههای کشور کمتر به این موضوع پرداخته اند. البته نمی توان اهمیت انتشار تعدادی کتاب در حوزه بازیابی اطلاعات یا تدوین اصطلاحنامه های موضوعی را نادیده گرفت. در قلمرو نمایه سازی نیز کتاب های قابل توجهی تألیف و ترجمه شده است. نکاتی که در بالا به آنها اشاره شد، به طور پراکنده در این آثار مورد بحث و مطالعه قرار گرفته اند. به خاطر اهمیت این نوع موضوعات در حیطه بازیابی اطلاعات بود که کتاب **معناشناسی و بازیابی اطلاعات**^۱ تهیه و در دسترس علاقه مندان به این موضوع قرار گرفت. به علاوه، توجه گروه های آموزشی کتابداری و اطلاع رسانی نیز به اهمیت این موضوع جلب شده و برای نخستین بار گروه کتابداری و اطلاع رسانی دانشگاه فردوسی مشهد، پژوهشی را به صورت پایان نامه در این قلمرو در دوره دکتری در دست اجرا دارد. مرکز منطقه ای اطلاع رسانی علوم و فناوری نیز در گروه پژوهشی زبانشناسی رایانه ای مطالعه ای را در حوزه ریشه ساز فارسی آغاز کرده و انتظار می رود نتیجه آن به عنوان اولین تجربه به زودی در دسترس کتابخانه ها و مراکز اطلاع رسانی قرار گیرد.

مطالبی که در فصول مختلف این کتاب در پیش روی دارید گزیده ای است آگاهانه از مطالعاتی که در مراکز علمی جهان در حیطه زبان فارسی، اسپانیولی و انگلیسی صورت گرفته است.

اطمینان داریم آنچه در این اثر آورده شده است کامل نبوده و صرفاً دریچه ای را در حوزه پردازش زبان طبیعی و بازیابی اطلاعات به روی پژوهشگران علوم کتابداری و اطلاع رسانی می گشاید.

همه انگیزه با رشد این موضوع فقط در صورتی امکان پذیر است که بتوان به بهترین نحو از منابع موجود در آن استفاده نمود.

دکتر جعفر مهران

استاد دانشگاه شیراز

۱. ترجمه جعفر مهران و محمد رضا فلاحتی. مشهد: انتشارات کتابخانه رایانه ای و مرکز منطقه ای اطلاع رسانی علوم و فناوری، ۱۳۸۴