

فراداده‌ها،
موتورهای کاوش وب
و میانکنش پذیری آنها

دکتر سید مهدی طاهری

(عضو هیئت علمی دانشگاه علامه طباطبائی)

دکتر نجلا حریری

(عضو هیئت علمی دانشگاه آزاد اسلامی، واحد علوم و تحقیقات تهران)

دکتر رحمت‌الله فتاحی

(عضو هیئت علمی دانشگاه فردوسی مشهد)



طاهری، سید مهدی، ۱۳۵۷ -
فراداده‌ها، موتورهای کلوش وب و میانکنش‌پذیری آنها
سید مهدی طاهری، نجلا حریری و رحمت‌الله فتاحی
تهران: کتابدار: انجمن بین‌المللی سازماندهی
دانش (شاخه ایران)، ۱۳۹۴
۲۰۱ ص: مصور، جدول (کتابدار: ۲۸۷)
فهرست‌نویسی بر اساس اطلاعات فیپا
ص. ع. به انگلیسی: Seyed Mahdi Taheri et al.
Web search engines and their interoperability

واژه‌نامه

کتابنامه

نمایه

موضوع: ۱. فراداده‌ها ۲. وب - موتورهای جستجو
الف. حریری، نجلا، ۱۳۴۰- نویسنده همکار ب. فتاحی،
رحمت‌الله، ۱۳۳۰- نویسنده همکار ج. عنوان
رده کنگره: ۱۳۹۴ ۴ط۲ / ۶۶۶/۷ Z
رده دیویی: ۰۲۵/۳
شماره کتابشناسی ملی: ۳۸۲۷۰۲۹



www.ketabdar.org
mohseni@ketabdar.org

عنوان: فراداده‌ها، موتورهای کلوش وب: میانکنش‌پذیری آنها
مؤلفان: دکتر سیدمهدی طاهری
دکتر نجلا حریری و دکتر رحمت‌الله فتاحی
مجموعه کتابدار، ۲۸۷

نوبت چاپ: اول، ۱۳۹۴ / شمارگان: ۱۰۰۰ نسخه

قیمت: ۱۴۰۰۰ تومان

شابک: ۹۷۸-۶۰۰-۲۴۱-۰۹۴-۸

تلفن: ۶۶۴۱۵۵۶۴ / ۶۶۴۱۵۵۶۱

دورنگار: ۸۸۹۷۹۵۶۰ / همراه: ۰۹۱۲۱۸۸۳۵۸۱

فهرست مطالب

۹	پیشگفتار
۱۱	مقدمه
۱۲	مفهوم سازماندهی دانش
۱۳	نظام‌های سازماندهی دانش
۱۳	ویژگی‌های مشترک نظام‌های سازماندهی دانش
۱۴	انواع نظام‌های سازماندهی دانش
۱۴	فهرست اصطلاحات
۱۵	رده‌بندی‌ها و طبقه‌بندی‌های
۱۵	فهرست‌های روابط
۱۶	فرا داده (کلیات)
۱۷	جایگاه فراداده در سازماندهی دانش
۱۸	مروری بر مفهوم فراداده
۲۰	تاریخچه شکل‌گیری اصطلاح و طرح‌های فراداده‌ای
۲۱	ضرورت و اهمیت استفاده از فراداده
۲۲	شیوه‌های پیوند فراداده‌ها با اشیای محتوایی
۲۳	کارکردهای فراداده
۲۴	مزایا و کاربردهای فراداده
۲۵	استانداردهای فراداده‌ای
۲۶	برخی نکات کلیدی در استفاده از فراداده
۲۸	طرح فراداده‌ای هسته دوبلین
۲۸	پدیده هسته دوبلین
۲۹	تاریخچه
۳۰	اهداف استاندارد فراداده‌ای هسته دوبلین
۳۱	واژگان کنترل شده و ارزش‌های عناصر هسته دوبلین
۳۲	مجموعه عناصر فراداده‌ای هسته دوبلین
۳۲	توضیحگرهای هسته دوبلین
۳۲	ضرورت استفاده از توضیحگرها
۳۳	انواع توضیحگرهای هسته دوبلین

- ۳۴ نرم افزارها و ابزارهای هسته دویلین
- ۳۵ بستر نحوی پیاده‌سازی پیشینه‌های فراداده‌ای هسته دویلین
- ۳۷ وضعیت کنونی طرح فراداده‌ای هسته دویلین
- ۳۸ قالب فراداده‌ای مارک ۲۱، و تمهیدات آن برای سازماندهی اشیای
- ۴۰ فیلد ۸۵۶
- ۴۱ قالب فراداده‌ای مارک ۲۱ در بستر زبان نشانه گذاری گسترش پذیر
- ۴۱ ویژگی‌ها و کاربردهای مارک در قالب ایکس.ام.ال.
- ۴۱ کاربردها
- ۴۱ مزایای مارک در قالب ایکس.ام.ال.
- ۴۲ محدودیت‌های مارک در قالب ایکس.ام.ال.
- ۴۲ تبدیل مارک ۲۱ (۲۷۰۹) به/از مارک در قالب ایکس.ام.ال.
- ۴۲ گذرگاه مارک در قالب ایکس.ام.ال.
- ۴۳ ملاحظات طراحی مارک در قالب ایکس.ام.ال.
- ۴۳ فرامای ساده و انعطاف پذیر مارک در قالب ایکس.ام.ال.
- ۴۳ تبدیل بدون آسیب مارک به مارک در قالب ایکس.ام.ال.
- ۴۳ قابلیت تبدیل مارک در قالب ایکس.ام.ال. به مارک
- ۴۳ ارائه داده‌ها
- ۴۴ ویرایش مارک
- ۴۴ تبدیل داده‌ها
- ۴۴ اعتبارسنجی داده‌های مارک
- ۴۴ گسترش پذیری
- ۴۵ طرح فراداده توصیف شیء (MODS)
- ۴۹ طرح فراداده‌ای توصیف مستند (MADS)
- ۵۰ بستر نحوی پیشینه‌های فراداده‌ای
- ۵۱ زبان نشانه گذاری فرامتن (HTML)
- ۵۲ برچسب `<pre>`
- ۵۳ زبان نشانه گذاری گسترش پذیر (XML)
- ۵۴ کاربردهای زبان نشانه گذاری گسترش پذیر
- ۵۵ استانداردهای خانواده بزرگ ایکس.ام.ال.
- ۵۶ آشنایی با مفاهیم، فناوری‌ها، و پایگاه‌های داده‌ای مبتنی بر
- ۵۶ تعریف زبان های ایکس.ام.ال.
- ۵۷ نحوه انتشار اشیای محتوایی ایکس.ام.ال.
- ۵۸ ایکس.ام.ال.، قالب انتقال داده ها
- ۵۹ زبان‌های تعریف سبک ایکس.ام.ال.
- ۵۹ شیوه نامه آبخاری
- ۵۹ زبان شیوه‌ای گسترش پذیر

۶۰	نحوه عملکرد شیوه نامه
۶۱	نحوه پردازش اشیای محتوایی مبتنی بر ایکس.ام.ال.
۶۲	تجزیه کننده
۶۳	ساختار یک شیء محتوایی مبتنی بر ایکس.ام.ال.
۷۱	مفاهیم خوش شکل و معتبر
۷۱	ویژگی های یک شیء محتوایی خوش شکل
۷۱	فرانمای زبان نشانه گذاری گسترش پذیر
۷۲	اعتبارسنجی اشیای محتوایی مبتنی بر ایکس.ام.ال.
۷۳	فناوری مسیر یاب ایکس.ام.ال.
۷۳	فناوری زبان درخواست ایکس.ام.ال.
۷۴	ویژگی های زبان درخواست ایکس.ام.ال.
۷۴	کاربردها
۷۵	پایگاه های داده ای مبتنی بر ایکس.ام.ال.
۷۵	تعریف پایگاه های داده ای مبتنی بر ایکس.ام.ال.
۷۵	پایگاه های با قابلیت پشتیبانی از ایکس.ام.ال.
۷۶	پایگاه های بومی ایکس.ام.ال.
۷۷	بازایی داده های ذخیره شده در پایگاه های داده ای مبتنی بر
۷۷	نمایه سازی اشیای محتوایی مبتنی بر ایکس.ام.ال.
۷۸	رویکرد مبتنی بر پایگاه داده ای
۷۹	رویکرد مبتنی بر بازایی اطلاعات
۸۰	رویکرد دو وجهی
۸۲	روش (یا راهبرد) جزیره داده ها
۸۴	موتورهای کاوش وب
۸۵	انواع ابزارهای کاوش وب
۸۵	موتورهای کاوش مبتنی بر پیمایش
۸۶	راهنماهای تقویت شده توسط انسان یا نمایه های دستی
۸۶	«ابزارهای کاوش دو وجهی» یا نتایج ترکیبی
۸۶	فراموتورهای کاوش
۸۷	تاریخچه موتورهای کاوش
۸۹	ساختار موتورهای کاوش
۹۰	روش های نمایه سازی موتورهای کاوش
۹۱	ساختارهای داده ای مربوط به نمایه
۹۲	نمایه های مقلوب
۹۳	ترکیب نمایه ای
۹۴	نمایه پیشین
۹۴	تجزیه مدرک

۹۵	تعیین کلمات رمزی
۹۵	شناخت زبان
۹۶	تحلیل قالب
۹۷	تشخیص بخش‌های مختلف اشیای محتوایی
۹۸	نمایه‌سازی فرابرچسب‌ها
۹۹	الگوی نمایه‌سازی موتورهای کاوش
۹۹	الگوی متمرکز
۱۰۰	الگوی توزیعی
۱۰۲	معرفی مهم‌ترین موتورهای کاوش موجود در وب
۱۰۲	گوگل
۱۰۴	ياهو
۱۰۴	شرکت ياهو
۱۰۴	راهنمای موضوعی و موتور کاوش ياهو
۱۰۵	ائتلاف موتور کاوش ياهو و موتور کاوش بینگ
۱۰۶	مروری بر پیشرفته‌ترین روش‌های معرفی وب سایت‌ها به
۱۰۶	معرفی مستقیم وب سایت‌ها
۱۰۹	امکان نقشه سایت مبتنی بر ایکس.ام.ال.
۱۱۲	تایید و تعیین اعتبار وب سایت‌های معرفی شده به موتورهای کاوش
۱۱۲	گوگل
۱۱۳	ياهو
۱۱۵	سخن پایانی
۱۱۵	تحلیلی بر مطالعات انجام شده در حوزه میانکنش‌پذیری
۱۳۰	منابع فارسی
۱۳۷	پیوست الف. فهرست عناصر دو سطح ساده و ویژه طرح فراداده‌ای هسته دویلین به همراه
۱۴۱	پیوست ب. عناصر، عناصر فرعی، و خصیصه‌های طرح فراداده‌ای توصیف شیء (نسخه ۳.۵)
۱۴۹	پیوست ج. فهرست برچسب‌ها و فرابرچسب‌های زبان نشانه‌گذاری فرامتن
۱۵۷	پیوست د. نحوه ایجاد پایگاه داده‌های مبتنی بر ایکس.ام.ال. در محیط
۱۷۱	واژه‌نامه انگلیسی - فارسی
۱۸۵	واژه‌نامه فارسی - انگلیسی
۱۹۹	نمایه موضوعی

پیشگفتار

بخش عمده محتوای کتاب پیش‌رو، برگرفته از رساله دکترای نگارنده است که در سال ۱۳۹۱ به پایان رسید. در این رساله، یکی از مهم‌ترین مسائل حوزه ذخیره و بازیابی اطلاعات و دانش (یعنی میانکنش‌پذیری نظام‌های فراداده‌ای و موتورهای کاوش وب) مطالعه و بررسی شد که بر بهبود دسترسی‌پذیری اشیای محتوایی وب از طریق ابزارهای عمومی کاوش وب تأکید دارد. انگیزه نگارش کتاب، اهمیت موضوع، مفید و کاربردی بودن محتوای اثر برای بهبود تلاش‌های کشور در زمینه میانکنش‌پذیری ابزارهای یاد شده، و توصیه اساتید هدایت‌کننده و داور رساله، و دیگر همکاران بوده است. مطالب کتاب به دلیل ماهیت میان‌رشته‌ای برای هر دو گروه حرفه‌مندان علم اطلاعات و علم رایانه مناسب خواهد بود. پوشش موضوعی کتاب و پرداختن عمیق و نسبتاً مفصل به موضوعات فراداده، موتورهای کاوش وب، و فرایند میانکنش‌پذیری آنها، قابلیت استفاده از کتاب به عنوان منبع درسی یا کمک درسی را فراهم نموده است. معرفی اثر برای واحدهای درسی سازماندهی منابع اطلاعاتی به ویژه سازماندهی رایانه‌ای اطلاعات در مقطع کارشناسی، بازیابی اطلاعات (نمایش‌سازی پیشین) به ویژه برای تبیین نمایه‌سازی وب یا موتورهای کاوش وب، ذخیره و بازیابی اطلاعات، ارزشیابی نظام‌ها و خدمات اطلاعاتی، و طراحی و توسعه کتابخانه‌های دیجیتال در مقطع کارشناسی ارشد، و واحدهای درسی نظریه‌های سازماندهی دانش، تحلیل متن و دانش‌کاوی، وب معنایی و هستی‌شناسی، و معماری اطلاعات در مقطع دکترای علم اطلاعات و دانش‌شناسی، و نیز واحدهای درسی مشابه در علم رایانه مفید به نظر می‌رسد. همچنین کاربرمداری و جامعیت نسبی توصیه‌های عملی ارائه شده در کتاب برای طراحان نظام‌های مبتنی بر وب و فهرست‌های رایانه‌ای راهگشا خواهد بود. برای حفظ کاربردپذیری و گسترش‌پذیری چشمگیر مطالب، روزآمدسازی‌های لازم در بخش‌های گوناگون اثر صورت گرفته است. عدم قدردانی از بندگان، ناسپاسی از خداوند متعال خواهد بود، بدینوسیله مراتب سپاس و تشکر خود را خدمت اساتید فرهیخته، جناب آقای دکتر سید رحمت‌الله فتاحی، سرکار خانم

دکتر نجلا حریری، و همکاران پرتلاش و فهیم، سرکار خانم‌ها اشرف فدائی و نعیمه برجیان (کارشناسان ارشد کتابخانه مرکزی دانشگاه علامه طباطبائی^{۲۰})، و به ویژه مدیر فرزانه نشر کتابدار همکار ارجمند جناب آقای حمید محسنی که نگارنده را برای انتشار کتاب تشویق و حمایت نمودند، ابراز می‌دارم. امید است محتوای کتاب برای متخصصان گرامی علوم اطلاعات و رایانه مفید و کاربردی ارزیابی شود (انشاءالله).

دکتر سید مهدی طاهری

عضو هیئت علمی دانشگاه علامه طباطبائی

اردیبهشت ۱۳۹۴

در حال حاضر شبکه جهانی وب بزرگترین محیط ذخیره و بازیابی اطلاعات است. دسترسی به اطلاعات این محیط نیز مانند محیط‌های سنتی نیازمند سازماندهی است. با این تفاوت که حجم گسترده اطلاعات و ویژگی‌های فنی وب، نظام‌ها و ابزارهای ویژه‌ای را برای سازماندهی محتوا می‌طلبد. فراداده، پاسخ حوزه سازماندهی دانش به مسئله دسترسی به محتوای محیط وب است. ابزار دیگر، موتورهای کاوش است که اقدامی فناورانه از سوی متخصصان حوزه فناوری اطلاعات و ارتباطات برای حل این مسئله است. با توجه به این که موتورهای کاوش بر اساس فرایند نمایه‌سازی، اطلاعات وب را جستجوپذیر می‌نمایند، به حوزه سازماندهی دانش نیز مرتبط می‌شوند. با این تفاوت که پیشینه‌های فراداده‌ای بیشتر توسط عوامل انسانی، و محتوای جستجوپذیر از طریق موتورهای کاوش بیشتر توسط عوامل ماشینی تولید می‌گردد. بنابراین، هر دو ابزار از طریق نظم بخشیدن و ساماندهی اطلاعات وب، دسترسی به این اطلاعات را تسریع و تسهیل می‌کنند.

ماهیت فنی و رسانه‌ای وب رویکردهای جدیدی را در حوزه سازماندهی پدید آورده است. پاره‌های اطلاعاتی که بیشتر به صورت مجموعه در یک شیء محتوایی ذخیره می‌شدند، در محیط وب به صورت مستقل ذخیره می‌شوند (دانه دانه)^۱، و برای دسترس‌پذیری نیاز به توصیف جداگانه دارند. تفکیک محتوا و تبدیل آنها به اشیای محتوایی^۲ مستقل (در عین وابستگی به شیء مادر) از یک سو، و قابلیت‌های منحصربه‌فرد آن یعنی فرایوند و چندرسانه‌ای از سوی دیگر، امکان برقراری روابط پیشین (با شیء مادر) و پسین (با شیء دیگر) میان اشیای محتوای وب را فراهم نموده است. از آنجا که تعامل میان اطلاعات به تولید دانش می‌انجامد، می‌توان گفت محیط وب به سمت تولید دانش حرکت نموده است. شبکه‌ای بودن بستر وب نیز ماهیت رابطه-مدار آن را نشان می‌دهد. از این رو، توصیف و سازماندهی دقیق محتوای وب (توصیف اشیای محتوایی و کشف و ثبت روابط میان آنها) بر پایه ماهیت رابطه-مدار

^۱ Granularity

^۲ Content objects

موجودیت‌هایی که داده‌ها یا اطلاعات را در بر می‌گیرند. یک شیء محتوایی می‌تواند از اشیای محتوایی دیگر ساخته شود. برای مثال مجله یک شیء محتوایی است که از مجموعه مقاله‌هایی درست شده است که هر کدام آنها یک شیء محتوایی مستقل است. متن، تصویر و شکل یک مقاله نیز می‌تواند یک شیء محتوایی باشد. نقشه، نقاشی، عکس، و دیگر اشیای غیرمتنی نیز شیء محتوایی محسوب می‌شوند. حتا اطلاعات کتابشناختی یا به اصطلاح فراداده‌های مربوط به یک شیء محتوایی نیز خودشان اشیای محتوایی می‌باشند (سازمان ملی استانداردهای اطلاعاتی، ۲۰۰۵). بر اساس این رویکرد، اشیای محتوایی می‌توانند ساختار سلسله مراتبی داشته باشند.

آن، سازماندهی دانش، و نه اطلاعات، محسوب می‌شود. وجود عناصر رابطه-مدار در طرح‌های فراداده‌ای، و بازیابی صفحات (اشیای محتوایی) مشابه^۱ توسط موتورهای کاوش، هم‌راستایی ابزارهای سازماندهی وب به ماهیت آن را بیان می‌کند. تعامل نظام‌های فراداده‌ای و موتورهای کاوش به دسترس‌پذیری محتوای وب خواهد افزود. و نمایه‌سازی و پیدانمایی^۲ نام‌های برچسب عناصر^۳ فراداده‌ای، سطح بالای میانکنش‌پذیری خواهد بود. در همین راستا، نخست به مفهوم سازماندهی دانش، و نظام‌های سازماندهی دانش و انواع آن پرداخته می‌شود. سپس فراداده به عنوان ابزار سازماندهی دانش وب، و بستر نحوی آن که مبنای استخراج ارزش‌ها و برچسب‌ها توسط نرم‌افزارهای خزنده-نمایه‌ساز موتورهای کاوش وب است مطرح می‌شود. در پایان، موتورهای کاوش با تاکید بر روش‌های نمایه‌سازی و معرفی وب‌سایت‌ها به عنوان رسانه‌های اشیای محتوایی به موتورهای کاوش برای فراهم نمودن بستر میانکنش‌پذیری توصیف می‌گردند.

مفهوم سازماندهی دانش

سازماندهی دانش یکی از مراحل و رویکردهای مهم در فرایند مدیریت دانش و دسترس‌پذیر ساختن منابع دانش است. این فرایند با توصیف و بازنمایی یکپارچه ویژگی‌های مادی و فکری منابع دانش سروکار دارد، و در ذات خود با مقوله‌هایی چون جامعه‌شناسی، جامعه‌گرایی، شناخت، نظم فکری، و نظیر آن مرتبط است. به بیان دیگر، به‌طور مستقیم و غیرمستقیم با ذهن افراد و جامعه در ارتباط است. بازتاب فکری افراد (به‌صورت فردی) و جامعه (به‌صورت کل) یعنی منابع مکتوب و مدون دانش بشری، شناسایی می‌شود، تهیه می‌شود، سازماندهی می‌شود و اشاعه داده می‌شود تا برای استفاده و دانش‌افزایی سایر افراد جامعه قابل شناسایی و دسترس‌پذیر شود.

^۱ Similar pages

^۲ Visibility

نمایش اشیای محتوایی موجود در پایگاه داده‌ای موتورهای کاوش (نمایه شده توسط موتورهای کاوش) در برابر درخواست‌های کاربران در فهرست نتایج جستجو (سولوان، ۲۰۱۱). به عبارت دیگر، هنگامی که شیء نمایه‌شده توسط موتور کاوش بر اساس درخواست کاربران، بازیابی و در نتایج جستجو نشان داده شود، به سطح پیدانمایی رسیده است. تمامی موتورهای کاوش بر این مطلب تاکید می‌کنند که ممکن است شینی توسط آنها نمایه شود، اما به دلایل گوناگون، بازیابی نشود، و در نتایج جستجو نباشد.

^۳ Element Tag names

هر عنصر فراداده‌ای در قالب زبان‌های نشانه‌گذاری دارای دو بخش اصلی است: ۱. نام برچسب؛ ارزش (محتوا). در ذیل نمونه‌ای از عنصر عنوان طرح فراداده‌ای هسته دویلین ارائه شده است. در این عنصر فراداده‌ای <title>ایران زمین</title>، "<title></title>" نام‌های برچسب عنصر، و "ایران زمین" ارزش (محتوای) آن می‌باشد.