تحقیقات کتابداری و اطلاع‌رسانی دانشگاهی،
دورهٔ ۴۹، شمارهٔ ۲، تابستان ۱۳۹٤، ص ۲۳۹-۲٥٤

Academic Librarianship and Information Research
Vol. 49, No. 2, Summer 2015, pp. 239-254

# سنجش میزان تکرار اطلاعات در بازیابی اطلاعات فارسی از وب با مقایسهٔ موتورهای کاوش عمومی

**سمیرا گوهری[۱]، لیلا مکتبی‌فرد[۲]، حمیدرضا جمالی مهموئی[۳]**

تاریخ دریافت: ۱۳۹۳/۱۰/۱۹ تاریخ پذیرش: ۱۳۹٤/٦/۲۳

## چکیده

**هدف:** پژوهش حاضر با هدف سنجش میزان تکرار اطلاعات بازیابی‌شدهٔ فارسی در وب و مقایسهٔ موتورهای کاوش از لحاظ میزان توانمندی در بازیابی کمترین تکرار در محتوای نتایج انجام گرفت.

**روش:** این پژوهش، از دستهٔ مطالعات توصیفی است که از روش‌های کمّی و آمار توصیفی بهره می‌گیرد و با توجه به مقایسه و ارزیابی موتورهای کاوش، در زمرهٔ تحقیقات ارزیابی نظام‌های بازیابی اطلاعات نیز محسوب می‌شود.

**یافته‌ها:** یافته‌ها نشان داد که موضوعاتی همچون حقوق، جغرافیا و ادبیات با بیش از ۷۰ درصد تکرار در هر موتور کاوش، بیشترین میزان را در بازیابی نتایج دارای محتوای تکراری دارند. موتور کاوش گوگل با بازیابی تنها ٤۲/۸ درصد کمترین میزان و موتور کاوش بینگ با بازیابی ۵۸/۳۳ درصد، بیشترین میزان را در بازیابی نتایج دارای محتوای تکراری داشتند.

**واژه‌های کلیدی:** اینترنت، بازیابی اطلاعات، تکرار اطلاعات، وب فارسی، موتورهای کاوش.

۱. دانشجوی کارشناسی ارشد علم اطلاعات و دانش‌شناسی، دانشگاه خوارزمی Gohari.samira68@yahoo.com
۲. استادیار علم اطلاعات و دانش‌شناسی، دانشگاه خوارزمی royamaktabi@gmail.com
۳. استادیار علم اطلاعات و دانش‌شناسی، دانشگاه خوارزمی h.jamali@gmail.com

## مقدمه

جست‌وجو و بازیابی اطلاعات از شبکهٔ جهان‌گستر وب، یکی از مباحث مهم یا به بیان بهتر یکی از چالش‌های موجود در پژوهش‌های اخیر حوزهٔ بازیابی اطلاعات محسوب می‌شود. شتاب فزایندهٔ رشد منابع گوناگون اطلاعاتی در این شبکه، ناهمگونی موضوعی و شکلی منابع، تنوع کاربران و نیازهای آنها، از جملهٔ مسائلی هستند که بازیابی اطلاعات از اینترنت را هر روز پیچیده‌تر می‌کنند (منصوریان، ۱۳۸۶). از اواسط دههٔ ۱۹۹۰، پژوهش دربارهٔ جست‌وجوی وب به حوزهٔ مهمی از اطلاعات مبدل شده است و به‌همین دلیل، بررسی نحوهٔ جست‌وجو و بازیابی در موتورها و ابرموتورهای کاوش نیز حوزهٔ مهمی از تحقیقات وبی به‌شمار می‌آید (بارایلان[1]، ۲۰۰۵). برای جست‌وجو و بازیابی اطلاعات از وب لازم بود تا ابزارهایی برای کاوش و دستیابی به اطلاعات مد نظر ایجاد شود. به این منظور ابزارهای کاوش همچون، موتور کاوش‌ها، ابرموتورهای کاوش و راهنماهای موضوعی ایجاد شدند، اما این کافی نبود و برای دستیابی به بهترین و مرتبط‌ترین اطلاعات، لازم بود تا این ابزارها بر اساس معیارهای از پیش تعیین‌شده‌ای ارزیابی شوند، تا بهترین آنها مشخص شود و کاربران بتوانند بر اساس آن، هر چه سریع‌تر به اطلاعات مورد نظرشان دست یابند.

در حال حاضر، برای ارزیابی ابزارهای جست‌وجوی اینترنت، معیارهای جدیدی ایجاد و استفاده می‌شود. یکی از این معیارها، ربط است که برای ارزیابی نظام‌های بازیابی اطلاعات و کارایی جست‌وجوهای انجام‌گرفته، به‌کار می‌رود و به‌وسیلهٔ دو ملاک جامعیت[2] و مانعیت[3] سنجیده می‌شود (شاکری، ۱۳۸۷). اما مشکل فقط وجود اطلاعات غیرمرتبط در نتایج جست‌وجو از وب نبود. طی جست‌وجوهای مختلف مشخص شد که بخش شایان توجهی از اطلاعات مرتبط یافت‌شده در فرایند بازیابی اطلاعات از وب، تکراری هستند، یعنی سایت‌ها و صفحاتی از وب که مطلبی را از سایت‌ها و صفحات دیگر گرفته‌اند و بدون استناد، با همان محتوا در سایت خود درج کرده‌اند.

سازماندهی، نمایه‌سازی، جست‌وجو و بازیابی اطلاعات در اینترنت نیز مشکلات گوناگونی به‌همراه دارد که اساساً به مسائلی مانند تحول در نوع و سطح اطلاعاتی که به آن افزوده می‌شود، شیوهٔ ورود اطلاعات و چگونگی دسترسی به آن مربوط خواهد بود. اکنون هر فرد یا سازمانی می‌تواند آثاری را به اینترنت اضافه کند و از این نظر هیچ کنترلی در کار نیست. این اطلاعات به‌صورت سازماندهی‌نشده به شرکت‌های متولی موتورهای جست‌وجو سپرده می‌شوند. برخلاف شیوهٔ مرسوم در کتابخانه‌ها و مراکز اسناد که منابع اطلاعاتی را بر اساس استانداردهای بین‌المللی سازماندهی می‌کنند، سایت‌ها و صفحات الکترونیکی در اینترنت بیشتر به طریق نمایه‌سازی

خودکار ذخیره می‌شوند. بازیافت فوق‌العاده زیاد منابع کم‌ربط یا نامرتبط و گاهی تکراری در اینترنت نیز به‌دلیل نامناسب بودن شیوه‌های نمایه‌سازی و سازماندهی اطلاعات در آن است (فتاحی، ۱۳۷۸).

اما علاوه بر تمامی مشکلات مذکور در زمینهٔ بازیابی اطلاعات، مسئله‌ای که در این پژوهش مطرح می‌شود این است که امروزه به‌دلیل نبود سازماندهی مناسب در محیط اینترنت، اطلاعات در زمینه‌های مختلف بدون اینکه استناد داده شوند و مؤلف و ناشر اصلی آنها معلوم باشد، نسخه‌برداری می‌شوند و نتیجه‌اش این است که کاربران هنگام بازیابی اطلاعات، با تعداد زیادی نتایج تکراری روبه‌رو می‌شوند. به‌عبارت دیگر می‌توان گفت که اطلاعات موجود بر روی وب بر اساس میزان ربطشان به سه دسته تقسیم می‌شوند: دسته‌ای از آنها کاملاً به موضوع ما نامرتبط هستند و امکان آگاهی‌بخشی ندارند؛ دسته‌ای کم‌ربط هستند و دستهٔ آخر که به موضوع مقاله مرتبط هستند، خود به دو دسته تقسیم می‌شوند: دست‌اول و تکراری. بیش از نیمی از این منابع بازگویی همان مطالب قبلی هستند که بدون استناددهی نسخه‌برداری می‌شوند و به اشغال فضای زیادی از اینترنت نیز منجر شده‌اند. بنابراین در این پژوهش قصد داریم که به این جنبه از مسئله، یعنی وجود تکرار در بازیابی اطلاعات فارسی در وب بپردازیم و با مقایسهٔ عملکرد چند موتور کاوش عمومی پراستفاده، کارآمدترین آنها را از لحاظ بازیابی کمترین اطلاعات تکراری در حوزه‌های موضوعی علوم انسانی معرفی کنیم، علاوه بر این تلاش خواهد شد که تا حد امکان با مشخص کردن خاستگاه اصلی اطلاعات، سایت‌های حاوی اطلاعات دست‌اول (منظور اطلاعاتی است که برای اولین بار توسط شخص یا سازمان خاصی تولید و در اینترنت عرضه شده‌اند) از سایت‌های حاوی اطلاعات تکراری و غیرمعتبر تمیز داده شود. امید است که نتایج این پژوهش برای پژوهشگران حوزهٔ علوم انسانی در شناسایی و دسترسی به منابع معتبر و استناددهی به منابع معتبر، مفید باشد.

## اهداف پژوهش

هدف اصلی پژوهش، سنجش میزان تکرار اطلاعات بازیابی‌شدهٔ فارسی در حوزه‌های موضوعی علوم انسانی در وب و معرفی بهترین موتور کاوش از لحاظ کمترین تکرار در محتوای نتایج بازیابی‌شده با شناسایی سایت‌های مرجع یا حاوی مطالب تکراری تا حد امکان است. با توجه به هدف اصلی پژوهش، اهداف فرعی آن عبارتند از:

- سنجش میزان تکرار اطلاعات بازیابی‌شدهٔ فارسی در حوزه‌های موضوعی علوم انسانی.
- مقایسهٔ حوزه‌های موضوعی علوم انسانی از لحاظ میزان تکرار اطلاعات بازیابی‌شدهٔ آنها در وب.

- مقایسهٔ توانمندی موتورهای کاوش از لحاظ کمترین میزان تکرار در بازیابی اطلاعات.
- مقایسهٔ میزان تکرار در نتایج حاصل از جست‌وجوی کلیدواژه‌ای و جمله‌ای.
- شناسایی و ارائهٔ فهرستی از سایت‌های مرجع در حوزهٔ موضوعی علوم انسانی تا حد امکان.
- شناسایی و ارائهٔ فهرستی از سایت‌های حاوی مطالب تکراری در حوزهٔ موضوعی علوم انسانی تا حد امکان.
- مقایسهٔ فراوانی سایت‌های مرجع و سایت‌های حاوی مطالب تکراری در میزان مشخصی از نتایج بازیابی‌شده از وب.

## مبانی نظری و پیشینهٔ پژوهش

### تعریف ذخیره و بازیابی اطلاعات

فعالیت‌هایی که برای تحلیل و سازماندهی مدارک و منابع صورت می‌گیرند، ذخیرهٔ اطلاعات و تلاش‌هایی که برای یافتن یک یا چند مدرک از میان انبوه مدارک ذخیره‌شده انجام می‌گیرند، بازیابی اطلاعات نام دارند. نظام‌هایی که این جریان‌ها در آنها روی می‌دهد، نظام‌های ذخیره و بازیابی اطلاعات خوانده می‌شوند (بهمن آبادی، ١٣٨۶، ج ١: ٨۵٣).

### بازیابی اطلاعات از وب

بازیابی اطلاعات در شبکهٔ جهانی وب، به‌وسیلهٔ ابزارهای جست‌وجو (موتورهای جست‌وجو و راهنماهای موضوعی) انجام می‌گیرد. ابزارهای جست‌وجو با نمایه کردن صفحات وب، پژوهشگر را از وجود اطلاعات در مکان‌های مختلف آگاه می‌کند. ابزارهای جست‌وجو اغلب در پاسخ به یک درخواست کاوش، ده‌ها یا صدها صفحهٔ وب را به‌صورت فهرست ارائه می‌دهند (سرچ انجین واچ، ٢٠٠۴).

ابزارهای کاوش اینترنت موضوع بحث‌های نظری و پژوهش‌های تجربی متعددی بوده‌اند، به‌گونه‌ای که بررسی و ارزیابی این ابزارها به‌منظور تشخیص کیفیت اطلاعات بازیابی‌شده توسط آنها، خود به یکی از مباحث مهم در حوزهٔ کتابداری و اطلاع‌رسانی تبدیل شده است (شاکری، ١٣٨٧).

ابزارهای کاوش اینترنت، پایگاه‌های اطلاعاتی جست‌وجو یا مرورپذیرند که با استفاده از آنها می‌توان به بخشی از اطلاعات موجود در اینترنت دست یافت. این ابزارها به دو شیوهٔ «جست‌وجوی

کلیدواژه‌ها» یا «مرور و انتخاب» پیوندهای فرامتنی کاربران را به‌سوی اطلاعات مورد نظر هدایت می‌کنند. ابزارهای کاوش اینترنت را می‌توان به دو نوع اصلی تقسیم کرد: راهنماهای موضوعی و موتورهای کاوش (کوشا، ۱۳۸۱).

**جست‌وجو و بازیابی در وب‌های نسل جدید**

وب در حال تکامل است و هر روزه فناوری‌های جدیدی در رابطه با وب اختراع یا ایجاد می‌شوند. اما تکامل وب به همین پیشرفت در فناوری محدود نشده است (رودکی، ۱۳۸۴). از آغاز پیدایش اینترنت و وب، مفاهیم جدیدی همچون وب۲، وب ۳، وب معنایی، وب پنهان و وب‌های نسل جدید مطرح شده‌اند که هر کدام از این مفاهیم به حوزهٔ خاصی از وب می‌پردازند. وب ۲، نسل جدید وب است که طراحی و معماری آن بر پایهٔ مشارکت‌ها، همکاری‌ها و تعاملات انسان‌ها و نرم‌افزار و دیگر عوامل هوشمند بنا نهاده شده‌اند. وب ۲ تمرکز را از تولیدکنندهٔ اطلاعات به مصرف‌کنندهٔ اطلاعات انتقال داده است. گفته می‌شود که وب ۲ فناوری جدیدی نیست، بلکه رویکرد جدیدی از وب محسوب می‌شود. شش چهرهٔ غالب را برای وب ۲ مطرح کرده‌اند که عبارتند از: وب مردمی‌تر، خودترسیمی، آموزش پیشرفته، پویایی داده‌ها، سرویس‌دهی، فضای سه‌بعدی (رودکی، ۱۳۸۴). نسل دیگری از وب، وب پنهان است که در سال ۱۹۹۴و توسط دکتر ژیل السورث و سال‌های بعد از ۱۹۹۴ توسط ماتیو مطرح و دنبال شد. تعریف کلاسیکی که از این نسل وب وجود دارد، عبارت است از: انواع منابع اطلاعاتی موجود در وب جهان‌گستر که به هر دلیل خارج از حوزهٔ بازیابی موتورهای کاوش عمومی قرار دارند. در بیان اهمیت وب پنهان می‌توان اهمیت آن را از دو بعد کمّی و کیفی به این‌ترتیب بررسی کرد: ازنظر کمّی، حجم اطلاعات نهفته در وب پنهان خیلی بیشتر از بخش سطحی یا آشکار آن است. از نظر کیفی، اطلاعاتی که در بخش‌های مختلف این مجموعه هستند، به‌ویژه منابع اطلاعاتی موجود در وب عمیق یا وب ملکی، معمولاً منابع مفید و ارزشمندی محسوب می‌شوند. بنابراین بحث در مورد این مقوله و به‌ویژه دربارهٔ بازیابی اطلاعات در این نسل جدید از وب حائز اهمیت است (منصوریان، ۱۳۸۲).

همان‌طور که در گذشته نیز اشاره شد، مبحث ذخیره و بازیابی اطلاعات، از موضوعات مورد توجه به‌ویژه در چند دههٔ اخیر بوده است و جنبه‌های مختلف تأثیرگذار بر فرایند جست‌وجو و بازیابی اطلاعات، از جمله تأثیر استفاده از اصطلاح‌نامه‌ها بر بهبود فرایند بازیابی، میزان دقت اطلاعات بازیابی‌شده، مقایسهٔ موتورهای کاوش از لحاظ توانمندی آنها در بازیابی اطلاعات مرتبط یا مقایسهٔ موتورها و ابرموتورهای کاوش از لحاظ میزان دقت و ... از مقوله‌هایی هستند که بسیار بر روی آنها کار شده است و در زیر به نمونه‌هایی از آنها اشاره می‌شود. اما تا آنجایی که

جست‌وجو شد، تاکنون پژوهشی بر روی این جنبه از موضوع صورت نگرفته است که بسیاری از اطلاعات بازیابی‌شده در حوزه‌های موضوعی مختلف تکراری هستند و در عمل هیچ‌گونه کاربردی ندارند. برخی پژوهش‌ها به‌طور ویژه به مقایسهٔ میزان دقت موتورهای کاوش عمومی پرداخته‌اند، از جملهٔ آنها می‌توان به پژوهش آزادی (۱۳۸۴) با عنوان «میزان دقت موتورهای کاوش وب در بازیابی اطلاعات کتابداری و اطلاع‌رسانی» اشاره کرد که به مقایسهٔ هفت موتور کاوش وب پرداخت. وی برای این کار از کلیدواژه‌های تخصصی حوزهٔ کتابداری و اطلاع‌رسانی در ده موضوع مختلف (که از مجلهٔ لایبرری ترندز[۴] گرفته شده بودند) استفاده کرد. موتورهای کاوش از نظر دقت به این‌ترتیب رتبه‌بندی شدند: اینفوسیک، هات بات، گوگل، آلتاویستا، اکسایت، لایکاس و وب کراولر. لیتون و سریواستاوا[۵] (۱۹۹۷) نیز با انتشار مقاله‌ای میزان مانعیت در بازیابی اطلاعات به‌وسیلهٔ پنج موتور جست‌وجوی آلتاویستا، هات‌بات، لایکاس، اکسایت و اینفوسیک را بررسی کردند. آنان جست‌وجوهای خود را با استفاده از ۱۵ موضوع متفاوت انجام دادند و پس از جست‌وجو، ۲۰ نتیجهٔ اول در هر موتور کاوش را برای تعیین ارتباط آنها با موضوعات مطرح‌شده بررسی کردند. آنها با تجزیه و تحلیل آماری به این نتیجه رسیدند که موتورهای کاوش آلتاویستا، اکسایت و اینفوسیک به‌ترتیب در رتبه‌بندی نتایج مرتبط، بهترین موتورهای جست‌وجو هستند.

بعضی از پژوهش‌ها، موتورهای جست‌وجو را از لحاظ معیارهایی همچون انعطاف‌پذیری، رابط جست‌وجوی قدرتمند و از نظر قابلیت‌های جست‌وجو یا سطح کارایی آنها ارزیابی کردند. چو و روزنتال[۶] (۱۹۹۶) در پژوهشی با عنوان «موتورهای کاوش خدمات شبکهٔ جهانی وب: مطالعهٔ مقایسه‌ای و روش ارزیابی» سه موتور جست‌وجوی وب شامل آلتاویستا، اکسایت و لایکاس را از نظر قابلیت‌های جست‌وجو (مثل منطق بولی، کوتاه‌سازی، فیلد جست‌وجو، کلمه و عبارت جست‌وجو) ارزیابی کردند و سطح کارایی آنها را (مثل دقت و زمان پاسخگویی) با استفاده از پرسش‌های جست‌وجوی نمونهٔ گرفته‌شده از سؤالات مرجع واقعی بررسی کردند و دریافتند که آلتاویستا نسبت به اکسایت و لایکاس از امکانات جست‌وجوی بهتری برخوردار است. در صورتی که میزان پوشش صفحات وب در موتور کاوش لایکاس، نسبت به دو موتور دیگر بیشتر بود. کورتوایز، بائر و استارک[۷] (۱۹۹۵)، در پژوهشی به بررسی و مقایسهٔ عملکرد و کارایی موتورهای کاوش سی‌یوآی، هاروست[۸]، اپن‌تکست[۹]، ورد وایدوب ورم، یاهو، و لایکاس و وب‌کراولر پرداختند. آنها برای این کار از سه پرسش جست‌وجوی نمونه استفاده کردند. نویسندگان به این نتیجه رسیدند که در بین آنها اپن‌تکست، از لحاظ انعطاف‌پذیری، رابط جست‌وجوی قدرتمند و

واکنش سریع بهترین بود. آنها همچنین به این نتیجه رسیدند که برای مبتدی‌ها، وب‌کراولر ساده‌ترین رابط را ارائه می‌دهد.

در این میان پژوهش‌هایی نیز به چشم می‌خورند که به مقایسهٔ موتورهای جست‌وجوی مربوط به کشورهای مختلف پرداخته‌اند. پژوهش ورونیس[10] (۲۰۰۶) که به مطالعهٔ تطبیقی سه موتور جست‌وجوی آمریکایی گوگل، یاهو و ام. اس.ان با سه موتور جست‌وجوی فرانسوی اکسالید[11]، ووئیلا[12] و دیر.کام[13] با استفاده از ۱۴ موضوع متفاوت پرداخته‌اند، نمونه‌ای از این پژوهش‌هاست که نتایج کار آنها نشان داد که موتور کاوش‌های آمریکایی گوگل و یاهو نسبت به رقبای خود عملکرد بهتری داشته‌اند. لواندوسکی[14] (۲۰۱۱) نیز در پژوهشی با عنوان «عملکرد بازیابی موتورهای جست‌وجو در مورد سؤالات مرجع[15]» سه موتور جست‌وجوی گوگل، یاهو و ام. اس. ان را با حدود ۹۰ درصد پاسخ درست در پاسخگویی به سؤالات مرجع، بهتر از سایر موتورها معرفی کرده است.

مرور پیشینهٔ پژوهش نشان می‌دهد که با وجود کثرت پژوهش‌ها در موضوع مقایسهٔ موتورهای کاوش در بازیابی اطلاعات، به‌تقریب تمام پژوهش‌ها یا به ارزیابی موتورها از لحاظ رابط کاربری و قابلیت‌های جست‌وجو یا از نظر میزان دقت و محاسبه جامعیت و مانعیت آنها پرداخته‌اند و بیشتر توجه پژوهشگران به این موضوع معطوف بوده است. ولی در این میان هیچ پژوهشی یافت نشد که به مقایسهٔ موتورها از لحاظ عملکرد آنها در ارائهٔ کمترین میزان تکرار در فهرست نتایج بازیابی‌شده بپردازد. پس در این پژوهش سعی شده است که تا حدودی به این مطلب پرداخته شود و با سنجش میزان تکرار در بازیابی نتایج، موتورکاوشی معرفی شود که تکرار کمتری در ارائهٔ اطلاعاتش دارد.

## روش‌شناسی پژوهش

پژوهش حاضر که پژوهشی کاربردی است برای سنجش میزان تکرار در موضوعات حوزهٔ علوم انسانی از روش‌های کمّی و آمار توصیفی بهره می‌گیرد و با توجه به مقایسه و ارزیابی موتورهای کاوش از لحاظ میزان توانمندی در بازیابی کمترین نتایج تکراری، در زمرهٔ تحقیقات ارزیابی نظام‌های بازیابی اطلاعات نیز محسوب می‌شود. همچنین برای شناسایی سایت‌های اصلی (مرجع) و تارنماهایی که مطالب سایر سایت‌ها را تکرار کرده‌اند، به‌نوعی از روش تحلیل محتوا استفاده می‌شود. بر این اساس پژوهش در سه مرحله انجام گرفت: در مرحلهٔ نخست، برای سنجش میزان تکرار اطلاعات موجود در وب از میان حوزه‌های موضوعی علوم انسانی، تمامی موضوعاتی که نشریات مربوط به آنها تا زمان انجام دادن این پژوهش (۱۳۹۲/۰۲/۰۱) در پایگاه مجلات تخصصی نور (نورمگز)[16] نمایه شده بود، مبنای عمل قرار گرفتند. به این ترتیب که در هر حوزهٔ موضوعی نشریه‌ای

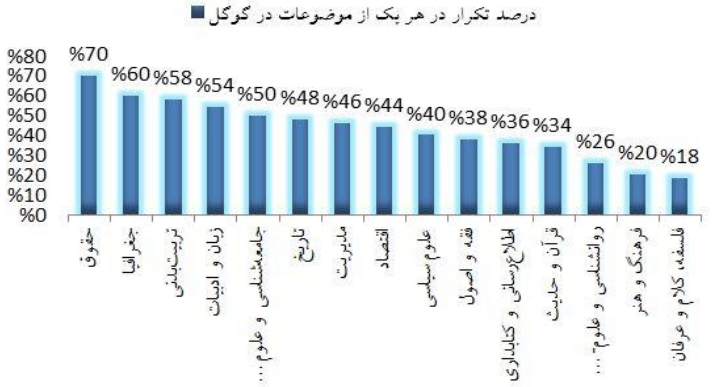
انتخاب شد که دارای رتبهٔ علمی - پژوهشی (تنها در یک مورد علمی - ترویجی) بوده و در پایگاه استنادی علوم جهان اسلام[17] نمایه شده و زبان آن نشریه فارسی است. برای انتخاب کلیدواژه‌ها چکیدهٔ مقالات مربوط به شماره‌های آخرین دوره (١٣٩١) هر یک از آن نشریات بررسی شد و از بین عنوان، چکیده و کلیدواژگان ارائه‌شده در چکیدهٔ هر مقاله، برای هر حوزهٔ موضوعی پنج عبارت موضوعی یا کلیدواژهٔ تخصصی استخراج شدند. به این‌ترتیب در مجموع، ٧٥ کلیدواژهٔ موضوعی از داخل پایگاه نورمگز برای کاوش در موتورهای جست‌وجو انتخاب شد تا مبنای انتخاب کلیدواژه‌ها سلیقه‌ای نباشد و از روی منابع علمی تخصصی مربوط به هر موضوع انتخاب شود.

در مرحلهٔ دوم با مرور پیشینهٔ پژوهش و همین‌طور با استفاده از موتور کاوش‌های عمومی معرفی‌شده در سایت سرچ انجین واچ[18] تعداد سه موتور کاوش که به‌عنوان پراستفاده‌ترین موتورهای کاوش شناخته شده‌اند، جامعهٔ پژوهش و برای مقایسه و ارزیابی انتخاب شدند. این موتور کاوش‌ها عبارتند از: گوگل[19]، یاهو[20] و بینگ[21]. سپس کلیدواژه‌های مربوط به هر حوزهٔ موضوعی برای سنجش میزان تکرار و مقایسهٔ موتورها از لحاظ میزان توانمندی در بازیابی کمترین نتایج تکراری، به هر کدام از سه موتور کاوش داده شد و ٢٠ نتیجهٔ اول هر جست‌وجو بررسی شدند و در نهایت در مرحلهٔ سوم برای دستیابی به بیشترین نتایج تکراری و شناسایی سایت‌هایی که مطالب سایت‌های دیگر را تکرار کرده‌اند و همین‌طور سایت‌های مرجع (یعنی سایت‌هایی که مطالبشان کپی‌برداری از سایت‌های دیگر نبود) از جست‌وجوی جمله‌ای استفاده شد. به این‌ترتیب که با بررسی نتایج به‌دست‌آمده از جست‌وجوی کلیدواژه‌ای، یک جملهٔ کلیدی که به بهترین نحو گویای محتوای آن کلیدواژه باشد، انتخاب و جملات انتخابی بدون استفاده از علامت نقل قول، برای جست‌وجو به موتورهای کاوش داده شد. بررسی ١٠ نتیجهٔ اول جست‌وجو نشان داد که جست‌وجوی جمله‌ای نسبت به جست‌وجوی کلیدواژه‌ای، بیشترین تکرار را به‌دست می‌دهد. با مقایسهٔ محتوای سایت‌های حاوی مطالب تکراری (که در بعضی موارد به منبع اصلی مطالب کپی‌برداری‌شده، استناد داده بودند) و در مورد خبرگزاری‌ها از روی تاریخ انتشار خبر و گاهی با توجه به ذکر منبع خبر، فهرستی از سایت‌های حاوی مطالب تکراری و سایت‌های مرجع ارائه شد.

## یافته‌های پژوهش

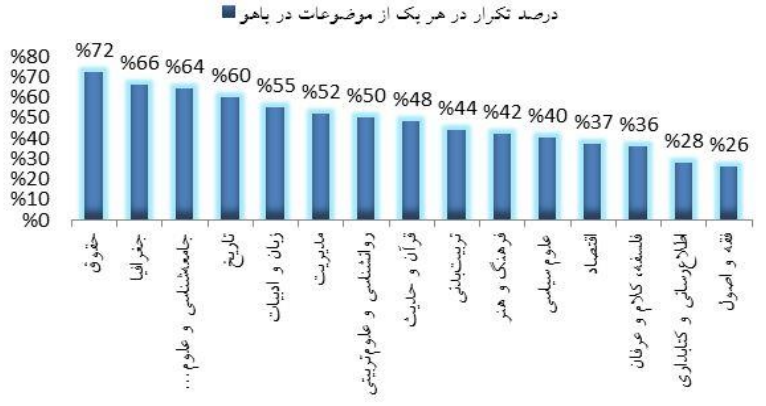
برای سنجش میزان تکرار در بازیابی نتایج مربوط به جست‌وجوی پانزده موضوع حوزهٔ علوم انسانی، کلیدواژه‌های تخصصی مربوط به هر موضوع، جداگانه در هر یک از موتورهای کاوش جست‌وجو شدند.

نمودار ۱ نشان‌دهندهٔ درصد فراوانی تکرار موضوعات به‌ترتیب از بیشترین تکرار تا کمترین تکرار در موتور کاوش گوگل است. مشخص شد که موضوع حقوق با اختصاص ۷۰ درصد بیشترین میزان تکرار را نسبت به سایر موضوعات و موضوع فلسفهٔ کلام و عرفان با اختصاص ۱۸ درصد، کمترین میزان تکرار را نسبت به سایر موضوعات در موتور کاوش گوگل داشته‌اند.



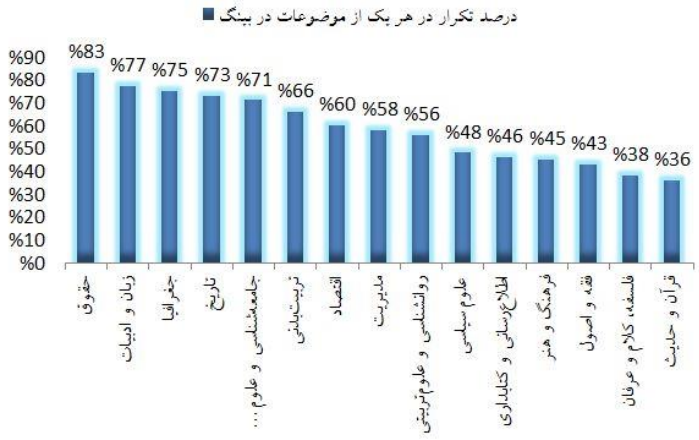

**نمودار ۱. مقایسهٔ میزان تکرار موضوعات در موتور گوگل**

نمودار ۲ نشان‌دهندهٔ درصد فراوانی تکرار موضوعات به ترتیب از بیشترین تکرار تا کمترین تکرار، در موتور کاوش یاهوست. مانند موتور گوگل، در این موتور نیز، موضوع حقوق با اختصاص ۷۲ درصد، بیشترین میزان تکرار را نسبت به سایر موضوعات دارد. اما در این موتور کمترین میزان تکرار مربوط به موضوع فقه و اصول با ۲۶ درصد است.




**نمودار ۲. مقایسهٔ میزان تکرار موضوعات در موتور یاهو**

نمودار ۳ نیز نشان‌دهندهٔ درصد فراوانی تکرار موضوعات به‌ترتیب از بیشترین تکرار تا کمترین تکرار، در موتور بینگ است. همچون دو موتور قبل، در موتور بینگ نیز، موضوع حقوق با اختصاص ۸۳ درصد، در صدر نمودار قرار دارد. اما در این موتور کمترین میزان تکرار مربوط به موضوع قرآن و حدیث با ۳۶ درصد است.

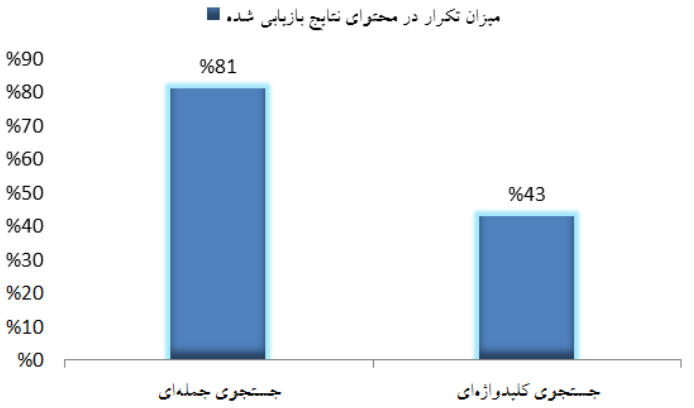



**نمودار ۳. مقایسهٔ میزان تکرار موضوعات در موتور بینگ**

نمودار ۴ به مقایسهٔ سه موتور کاوش گوگل، یاهو و بینگ در تمامی موضوعات بررسی‌شده، از لحاظ میزان توانمندیشان در بازیابی کمترین نتایج تکراری پرداخته است. همان‌طور که ملاحظه می‌شود، موتور کاوش گوگل با ۴۲/۸ درصد، کمترین تکرار را در بازیابی نتایج داشته و موتور کاوش بینگ با ۵۸/۳۳ درصد، بیشترین تکرار را نشان داده است.


**نمودار ٤. مقایسهٔ توانمندی سه موتور کاوش گوگل، یاهو و بینگ از لحاظ کمترین تکرار در بازیابی نتایج**

نتایج جست‌وجوی جمله‌ای، نشان داد که این نوع جست‌وجو در مقایسهٔ با جست‌وجوی کلیدواژه‌ای بیشترین میزان تکرار را در محتوای نتایج بازیابی‌شده ارائه می‌دهد. نمودار ۵ نتایج دو نوع جست‌وجو را مقایسه می‌کند.



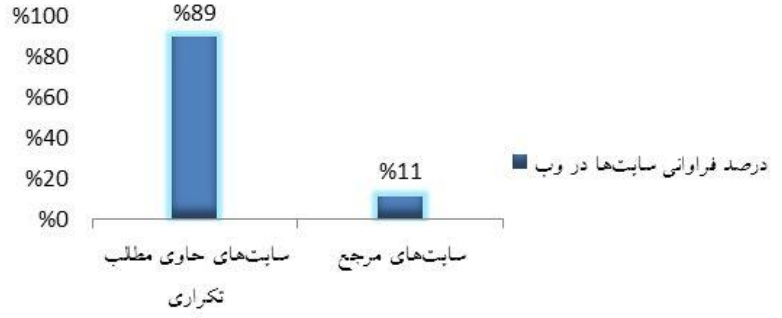

**نمودار ۵. مقایسهٔ میزان تکرار در محتوای نتایج جست‌وجوی جمله‌ای و کلیدواژه‌ای**

**جدول ۱. مقایسهٔ فراوانی سایت‌های مرجع و سایت‌های حاوی مطالب تکراری**

| مجموع نتایج حاصل از جست‌وجوی ۷۵ جملهٔ کلیدی و بررسی ۱۰ نتیجهٔ اول | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ۷۵۰ | | | | | | | | | | | |
| مجموع سایت‌های حاوی مطالب تکراری حاصل از جست‌وجوی جمله‌ای | | | | | | مجموع سایت‌های مرجع حاصل از جست‌وجوی جمله‌ای | | | | | |
| ۶۰۸ ٪۸۹ | | | | | | ٪۱۱۷۵ | | | | | |
| مجموع سایت‌های حاوی مطالب تکراری شناسایی‌شده | | | | | | مجموع سایت‌های مرجع شناسایی‌شده | | | | | |
| ۴۰۲ ٪۶۶ | | | | | | ۴۸ ٪۶۴ | | | | | |
| وبلاگ‌ها | خبرگزاری‌ها | | سایر سایت‌ها، پایگاه‌ها و انجمن‌ها | | | دایرةالمعارف‌های آنلاین | خبرگزاری‌های معتبر | | پایگاه‌ها و سایت‌های تخصصی | | |
| فراوانی | درصد فراوانی | فراوانی | درصد فراوانی | فراوانی | درصد فراوانی | فراوانی | درصد فراوانی | فراوانی | درصد فراوانی | فراوانی | درصد فراوانی |
| ۱۸۵ | ٪۴۶ | ۱۲۵ | ٪۳۱ | ۹۲ | ٪۲۳ | ۲۱ | ٪۴۴ | ۱۵ | ٪۳۱ | ۱۲ | ٪۲۵ |
| مجموع سایت‌های حاوی مطالب تکراری شناسایی‌نشده | | | | | | مجموع سایت‌های مرجع شناسایی‌نشده | | | | | |
| ۲۰۶ ٪۳۴ | | | | | | ٪۳۶۲۷ | | | | | |

در جدول ۱ مجموع نتایج حاصل از جست‌وجوی ۷۵ جملهٔ کلیدی و بررسی ۱۰ نتیجهٔ اول آن (که ۷۵۰ نتیجه می‌شود) ارائه شده است. از ۷۵۰ نتیجهٔ به‌دست‌آمده ۶۰۸ نتیجه یعنی ۸۹ درصد مربوط به سایت‌های حاوی مطالب تکراری بوده‌اند که با بررسی‌های انجام‌گرفته ۴۰۲ مورد معادل ۶۶ درصد از آن سایت‌ها شناسایی شدند و ۷۵ نتیجهٔ آن یعنی ۱۱ درصد به سایت‌های مرجع مربوط بوده‌اند که ۴۸ مورد آن، معادل ۳۴ درصد شناسایی و ارائه شد. در زمینهٔ سایت‌های حاوی مطالب تکراری، به‌ترتیب وبلاگ‌ها با فراوانی ۴۶ درصد، خبرگزاری‌ها با ۳۱ درصد و سایر سایت‌ها با ۲۳ درصد بیشترین تکرار را نشان دادند. در زمینهٔ سایت‌های مرجع نیز به‌ترتیب دایرةالمعارف‌هایی همچون ویکی‌پدیاها با ۴۴ درصد، سپس خبرگزاری‌های معتبر با ۳۱ درصد و در نهایت پایگاه‌ها و سایت‌های تخصصی با ۲۵ درصد، بیشترین فراوانی را داشته‌اند. جدول ۱ فراوانی و درصد فراوانی هر یک از سایت‌های مرجع و سایت‌های حاوی مطالب تکراری را نشان می‌دهد.

نمودار ۶ نشان‌دهندهٔ مقایسهٔ فراوانی سایت‌های حاوی مطالب تکراری و سایت‌های مرجع در میزان مشخصی از نتایج بازیابی‌شده از وب است. سایت‌های حاوی مطالب تکراری با ۸۹ درصد فراوانی، حجم زیادی از فضای وب را اشغال کرده‌اند و سایت‌های مرجع با ۱۱ درصد فراوانی در مقایسه با این سایت‌ها، فراوانی بسیار اندکی دارند.




**نمودار ٦. مقایسهٔ فراوانی سایت‌های حاوی مطالب تکراری و سایت‌های مرجع در وب**

## بحث و نتیجه‌گیری

با توجه به یافته‌های پژوهش و سنجش میزان تکرار اطلاعات در حوزه‌های موضوعی علوم انسانی از وب، آنچه به‌راحتی برداشت می‌شود، این است که حجم عظیمی از اطلاعات موجود در وب در زمینهٔ موضوعات حوزهٔ علوم انسانی (صرف‌نظر از تفاوت‌های اندکی که بین آنها وجود دارد) تکراری هستند و در عمل می‌توان گفت که تولید و ورود اطلاعات جدید در وب فارسی بسیار اندک است و فضای عظیمی از وب را اطلاعات تکراری اشغال کرده‌اند. در زمینهٔ تعداد بسیار

تکرار در برخی موضوع‌ها مانند حقوق، احتمال می‌رود که با توجه به اینکه حقوق از جمله حوزه‌هایی است که افراد به آن نیازمندند (و چون در جامعهٔ ما فرهنگ مراجعه به وکیل برای کسب اطلاعات حقوقی چندان رایج نیست) افراد ترجیح می‌دهند که از طریق رسانه‌هایی مانند اینترنت نیازهای اطلاعاتی خود را برطرف کنند. این تقاضا به‌دنبالش یک عرضه دارد که موجب انتشار این حجم از مطالب حقوقی در اینترنت شده است. با کمی دقت در مورد سایر موضوع‌های پرتکرار هم، همین اقبال عمومی به‌چشم می‌خورد. جغرافیا، تاریخ و ادبیات از جملهٔ حوزه‌هایی هستند که مردم معمولاً علاقه‌مندند در مورد آنها بیشتر بدانند. طبیعی است که اقبال عامهٔ مردم به موضوع‌هایی مانند مدیریت، کلام، حدیث و ... کمتر است. اگر امکان انجام دادن پژوهش دیگری برای سنجش میزان تکرار در سایر حوزه‌های موضوعی هم وجود داشته باشد، چه بسا نتایج حاکی از آن باشد که این حجم از تکرار در حوزه‌هایی همچون پزشکی (که یکی از حوزه‌هایی است که بیشتر مردم به اطلاعات آن حوزه، چه برای رفع نیازهای اطلاعاتی و چه برای ارتقای اطلاعات عمومی نیازمندند) بیش از سایر حوزه‌هاست. به‌عبارتی اگر آمار دقیقی از اطلاعات موجود در هر موضوع در محیط وب وجود داشته باشد، شاید بتوان گفت که این میزان از تکرار، در حقیقت تابعی از حجم اطلاعات موجود در این زمینه‌اند. اما طبیعی است که داشتن چنین آماری به‌تقریب محال خواهد بود.

موتور کاوش گوگل با بازیابی تنها ۴۲/۸ درصد تکرار در محتوای نتایج، نسبت به دو موتور بررسی‌شدهٔ دیگر (یاهو و بینگ)، بهترین عملکرد را در بازیابی نتایج با محتوای تکراری دارد. گاهی با انجام دادن چنین پژوهش‌هایی دلایل اقبال عمومی به یک سرویس خاص روشن‌تر می‌شود. نتایج حاصل از این پژوهش یک بار دیگر نشان داد که گوگل نسبت به سایر موتورهای جست‌وجو امتیازاتی دارد که آن را تا حدودی متمایز و پرمخاطب می‌کند (حتی برای کاربران غیرحرفه‌ای). پس از آن یاهو با ۴۸ درصد، کمترین میزان تکرار را داشت. موتور کاوش بینگ با ۵۸/۳۳ درصد نسبت به دو موتور دیگر، بیشترین میزان تکرار را دربازیابی نتایج نشان داد. نتایج این پژوهش با پژوهشی که ورونیس (۲۰۰۶) انجام داد و به بررسی تطبیقی شش موتور جست‌وجو پرداخت و عملکرد کلی آنها را در بازیابی بهتر نتایج ارزیابی کرد، همخوانی دارد. نتایج حاکی از آن بود که عملکرد گوگل و یاهو در بازیابی نتایج، بهتر از رقبای خود است. لواندوسکی (۲۰۱۱) نیز در پژوهشی عملکرد بازیابی موتورهای جست‌وجو را در پاسخ به سؤالات جست‌وجوشده ارزیابی کرد. یافته‌ها نشان داد که سه موتور جست‌وجوی گوگل و یاهو و ام. اس. ان با حدود ۹۰ درصد پاسخ درست به پرسش‌های جست‌وجو، بهترین عملکرد را داشته‌اند. می‌توان نتیجه گرفت

که عملکرد موتورهای کاوش گوگل و یاهو در بازیابی بهترین نتایج، چه از لحاظ میزان ربط و چه از جنبهٔ کمترین ریزش کاذب و کمترین تکرار در محتوای نتایج بازیابی‌شده، بهتر از سایر موتورهای کاوش است.

نتیجهٔ دیگری که از بررسی یافته‌های پژوهش حاصل شد، این بود که با انجام دادن جست‌وجوی جمله‌ای، میزان تکرار در محتوای نتایج بازیابی‌شده به نحو شایان ملاحظه‌ای افزایش می‌یابد، به‌طوری که میزان تکرار در نتایج حاصل از جست‌وجوی کلیدواژه‌ای ۴۲/۸ درصد و در جست‌وجوی جمله‌ای ۸۱ درصد بود. احتمال می‌رود که در جست‌وجوی جمله‌ای به‌علت تعداد زیاد کلمات مورد جست‌وجو، موتور کاوش سعی در یافتن مطالبی دارد که تمام کلمات را با هم داشته باشد و به بالاترین میزان ربط در نتایج دست یابد و این عمل منجر می‌شود که همهٔ مطالبی که عیناً حاوی آن جمله هستند بازیابی شوند و چه بسا که همگی آن نتایج یک مطلب باشند، ولی در سایت‌هایی با نشانی‌های گوناگون بازیابی شوند. با توجه به اینکه بیشتر کاربران مبتدی، جست‌وجوهای خود را اغلب در قالب جمله و به‌ویژه جملهٔ سؤالی مطرح می‌کنند، توصیه می‌شود که برای جلوگیری از دستیابی به نتایجی با محتوای تکراری، از جست‌وجوی کلیدواژه‌ای استفاده کنند.

خاستگاه اصلی اطلاعات بازیابی‌شده در وب یا به‌عبارتی سایت‌های مرجع را عمدتاً دایرةالمعارف‌هایی همچون ویکی‌پدیا، دانشنامه‌ها و پس از آن سایت‌های تخصصی (که در مورد هر حوزهٔ تخصصی وجود دارد) تشکیل می‌دهند. در مورد مطالب خبری، خبرگزاری‌های معتبری هچون خبرگزاری فارس و باشگاه خبرنگاران (بزرگ‌ترین خبرگزاری فارسی‌زبان دنیا) و ... خاستگاه اصلی مطالب خبری فارسی‌زبان هستند.

بیشترین میزان رونوشت‌برداری از مطالب سایت‌ها، به‌وسیلهٔ وبلاگ‌ها و خبرگزاری‌ها صورت می‌گیرد. به‌نظر می‌رسد قالب وبلاگ، همچنان جایگاه و کارکرد خود را به‌عنوان محملی برای روزنگاشت‌ها حفظ کرده است و با توجه به حجم عظیم مطالب تکراری و گاهی بی‌منبع و بی‌استنادی که در آنها یافت می‌شود، همچنان نمی‌توان وبلاگ را منبعی علمی به‌شمار آورد.

فراوانی سایت‌های مرجع در وب نسبت به سایت‌های حاوی مطالب تکراری به نحو شایان ملاحظه‌ای اندک است، به‌طوری که می‌توان گفت بیش از ۸۵ درصد از فضای وب را اطلاعات تکراری و زائد اشغال کرده‌اند و تولید اطلاعات فارسی دست‌اول و جدید در وب بسیار کم است.

## پی‌نوشت

1. Bar-Ilan
2. Recall
3. Precision
4. Trends Library
5. Leighton, Srivastavay
6. Chu, Rosental
7. Courtois, Baer, and Stark
8. Harvest
9. Open Text
10. Véronis
11. Exalead
12. Voilà
13. Dir.com
14. Lewandowski
15. Navigational queries
16. Noormag
17. ISC (Islamic World Science Citation Center)
18. Search Engine wath
19. Google
20. Yahoo
21. Bing

## منابع


۱. آزادی، قاسم (۱۳۸۴). میزان دقت موتورهای کاوش وب در بازیابی اطلاعات کتابداری و اطلاع‌رسانی. *فصلنامهٔ کتاب*. ۱۵ (۳)، ۱۱۱ - ۱۲۲.

۲. بهمن‌آبادی، علیرضا (۱۳۸۶). ذخیره و بازیابی اطلاعات. *دایرةالمعارف کتابداری و اطلاع‌رسانی*. ج ۱، ۸۵۳.

۳. رودکی، مهدی (۱۳۸۴). وب۲: موجی زودگذر یا آیندهٔ وب. *دنیای کامپیوتر و ارتباطات*. ۴۱.

۴. رودکی، مهدی (۱۳۸۴). وب ۲ چیست؟. *دنیای کامپیوتر و ارتباطات*. ۴۲.

۵. شاکری، صدیقه (۱۳۸۷). بررسی نسبت جامعیت و مانعیت ابزارهای کاوش اینترنت در بازیابی اطلاعات در حوزهٔ کتابداری و اطلاع‌رسانی. *مجلهٔ مطالعات ملی کتابداری و سازماندهی اطلاعات*. ۱۹ (۱)، ۱۷۷ - ۲۰۰.

۶. فتاحی، رحمت‌الله (۱۳۷۸). بلبشوی اینترنت: گفتاری پیرامون مشکلات سازماندهی، جست‌وجو، و بازیابی اطلاعات در وب جهان‌گستر. *فصلنامهٔ کتابداری و اطلاع‌رسانی*. ۲ (۲)، ۱ - ۲۲.

۷. کوشا، کیوان (۱۳۸۱). *ابزارهای کاوش اینترنت: اصول، مهارت‌ها و امکانات جست‌وجو در وب*. تهران: نشر کتابدار.

۸. منصوریان، یزدان (۱۳۸۲). مروری بر پژوهش‌های کاربرمدار در مطالعات بازیابی اطلاعات مبتنی بر وب. *نشریهٔ کتابداری و اطلاع‌رسانی*. ۶ (۳).

۹. منصوریان، یزدان (۱۳۸۶). عوامل مؤثر بر جست‌وجو بازیابی اطلاعات از شبکهٔ جهان‌گستر وب. *مجلهٔ الکترونیکی نشر کتابدار*.

10. Bar-Ilan, J. (2005). Comparing ranking of search result on the web search. *Information Processing & Management*, 41(6), 1511-1519.
11. Chu, H., & Rosenthal, M. (1996). *Search engines for the World Wide Web: a comparative study and evaluation methodology*. Retrieved Feb, 13, 2013, from http://www.asis.org/annual-96/ElectronicProceedings/chu.htm
12. Courtois, M., & Baer, W., & Stark, M. (1995*). Cool tools for searching the Web: A performance evaluation*. Online, 19(6), 14-32. Retrieved February 12, 2012, fromhttp://www.eric.ed.gov/ERICWebPortal/search/detailmini.jsp?_nfpb=true&_&ERICExtSearch_SearchValue_0=EJ515004&ERICExtSearch_SearchType_0=no&accno=EJ515004
13. Goodwin, D. (2013). Google Grabs More Search Market Share, Now at 67.5%. *Search Engine Watch*. Online, Retrieved June 6, 2013, from http://searchenginewatch.com/article/2255183/Google-Grabs-More-Search-Market-Share-Now-at-67.5
14. Leighton, H., & Srivastavay, J. (1997). *Precision among world wide web search services/search engines: Altavista, Excite, HotBot, Infoseek, Lycos*. Retrieved June 11, 2013, from http://www.winona.edu/library/webind2.htm.
15. Lewandowski, D. (2011). *The retrieval effectiveness of search engines on navigational*. Retrieved June 15, 2013, from http://eprints.rclis.org/17233/1/ASLIB2011_preprint.pdf
16. Véronis, J. (2006). *A comparative study of six search engines*. Retrieved June 11, 2012, from http://www.up.univmrs.fr/veronis/pdf/2006-comparative-study.pdf.