معیارهای ارزیابی ربط در نظام‌های بازیابی اطلاعات: دانسته‌ها و ندانسته‌ها

نجلا حریری[1] | فهیمه باب‌الحوائجی[2] | مهرداد فرزندی‌پور[3]
سمیه نادی راوندی[4]

۱. دانشیار؛ گروه علم اطلاعات و دانش‌شناسی؛ دانشگاه آزاد اسلامی؛ واحد علوم و تحقیقات تهران nadjlahariri@gmail.com

۲. دانشیار؛ گروه علم اطلاعات و دانش‌شناسی؛ دانشگاه آزاد اسلامی؛ واحد علوم و تحقیقات تهران f.babalhavaeji@gmail.com

۳. دانشیار؛ گروه فناوری اطلاعات سلامت؛ دانشگاه علوم پزشکی کاشان farzandipour_m@kaums.ac.ir

۴. [پدیدآور رابط] دانشجوی دکتری، گروه علم اطلاعات و دانش‌شناسی؛ دانشگاه آزاد اسلامی؛ واحد علوم تحقیقات تهران nadi_so@kaums.ac.ir

چکیده: ارزیابی نظام‌های بازیابی اطلاعات یکی از چالش‌های بزرگ متخصصان علم اطلاعات است؛ چون تعیین کارایی یک نظام به قضاوت در مورد ربط مدارک ارائه‌شده با نیاز اطلاعاتی کاربر بستگی دارد و این موضوع دارای پیچیدگی‌های خاصی است. از طرف دیگر، نظام‌های بازیابی وبی، بسیار متفاوت‌تر از نظام‌های بازیابی سنتی هستند. در محیط وب با نظام‌هایی با مجموعه‌های رتبه‌بندی‌شده و نیز با نظام‌هایی با مجموعه‌های رتبه‌بندی‌نشده مواجه هستیم که هر کدام معیار و مقیاس ارزیابی خود را دارند. به همین دلیل، سنجش ربط بر اساس نوع رتبه‌بندی نتایج در این نظام‌ها متفاوت است. در این مقاله معیارهای سنجش ربط در نظام‌های بازیابی اطلاعات به تفکیک نظام‌های رتبه‌بندی‌نشده (شامل دقت، بازیافت، معیار اِف) و نظام‌های رتبه‌بندی‌شده (شامل منحنی‌های دقت و بازیافت، دقت افزوده، میانگین دقت افزوده در نقطهٔ ۱۱، میانگین دقت متوسط، دقت در K، دقت R و منحنی ROC و افزایش تجمعی) تشریح شده و ضمن اشاره به مقیاس‌های ارزیابی ربط، یک مقیاس ارزیابی دقت ۴ درجه‌ای معرفی گردیده است که با استفاده از آن می‌توان به ارزیابی غیردودویی دقت در نظام بازیابی اطلاعات وب شامل دقت کامل، برتر، مفید، واقعی و در نهایت دقت متغیر پرداخت.

کلیدواژه‌ها: ارزیابی؛ بازیابی اطلاعات؛ دقت؛ معیار ارزیابی؛ مجموعه‌های رتبه‌بندی‌شده؛ مجموعه‌های رتبه‌بندی‌نشده

مقالهٔ پژوهشی

دریافت: ۱۳۹۲/۰۸/۱۲
پذیرش: ۱۳۹۲/۱۱/۱۹

دورهٔ ۳۰ شماره ۱
ص‌ص. ۱۹۹-۲۲۱

پژوهشنامهٔ پردازش و مدیریت اطلاعات
فصلنامه | علمی پژوهشی
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISA، ISC و Scopus
http://jipm.irandoc.ac.ir
پژوهشگاه علوم و فناوری اطلاعات ایران

۱. مقدمه

امروزه ارزیابی نظام‌های بازیابی اطلاعات بخش ضروری از تلاش‌های پژوهشگران علم اطلاعات شده است تا با استفاده از آن بتوانند میزان موفقیت و اثربخشی یک نظام را مشخص سازند. به گفتهٔ کارترت در هر ارزیابی دو سؤال اساسی مطرح است: نظام چقدر خوب است و چقدر بهتر و ارزشمندتر از نظام قبلی است؟ (Carterette 2008) ارزیابی به‌طور کلی، یک نیروی عمده در پژوهش، توسعه، و کاربردهای مربوط به نظام‌های بازیابی اطلاعات است (Saracevic 1995). نقش حیاتی ارزیابی در توسعهٔ ابزارهای بازیابی اطلاعات، گواه مناسبی در بهبود عملکرد این ابزارها و کیفیت نتایجی است که آنها ارائه می‌دهند (Bouramoul, Kholadi and Doan 2011). این ارزیابی می‌تواند از طریق دو رهیافت انجام شود: ارزیابی کاربرمدار و ارزیابی نظام‌مدار؛ که اولی از دید کاربر موضوع را بررسی می‌کند و دیگری سیستم را هدف قرار می‌دهد (Meadow 1992). اما صحبت از ارزیابی نظام بازیابی اطلاعات، همواره مفهوم ربط را به‌عنوان عمومی‌ترین معیار سنجش به ذهن متبادر می‌سازد که هر چند امری نسبی به نظر می‌رسد، تلاش برای افزایش آن، یعنی تطابق بین نیاز کاربر و پاسخ نظام‌های بازیابی همچنان ادامه دارد. با وجود اهمیت این مفهوم، هنوز هم تعریف، ماهیت، و ابعاد آن از سوی صاحب‌نظران به گونه‌های متفاوتی تفسیر و تأویل می‌گردد و دلیل، آن است که ربط از اساس دارای ماهیتی پویا، پیچیده و چندبعدی است (Schamber, Eisenberg and Nilan 1990). با این حال، سنجش این معیار، زمینه‌ساز سنجش اثربخشی یک نظام ذخیره و بازیابی اطلاعات به‌شمار می‌آید که به‌دلیل دخالت عوامل مختلف از جمله دانش کاربر، عمق نیاز، محیط بازیابی، و شناخت کاربر از نیاز خود و اطلاعات بازیابی‌شده (Hagglund 2004 نقل در حسن‌زاده ۱۳۸۷) دارای پیچیدگی‌های خاصی است؛ زیرا برای این نوع سنجش، عمومی‌ترین مقیاس‌ها، دقت و بازیافت است (Snášelet al 2009) که با وجود داشتن فرمول‌های ساده، به‌سادگی اندازه‌گیری نمی‌شوند. از طرف دیگر، با وجود استفادهٔ عمومی از آنها برای سنجش، تنها مقیاس‌های ارزیابی نیستند و در سطحی بالاتر همیشه مقیاس و معیار صحیح ارزیابی در هر نظامی محسوب نمی‌شوند؛ چرا که طبیعت پویای وب و امکانات و پیشرفت‌های فناوری بر نظام‌های بازیابی اطلاعات تأثیر فراوانی گذاشته است. از آنجا که نظام‌های وبی به دو دستهٔ نظام‌های رتبه‌بندی‌نشده و نظام‌های رتبه‌بندی‌شده تقسیم می‌شوند، ارزیابی دقت و بازیافت

هر کدام از این نظام‌ها دارای معیارهای متفاوتی است که در ایران کمتر به آن توجه شده است. جستجوی کلیدواژه‌های ربط، دقت، بازیافت، معیار اِف، معیارهای ارزیابی، منحنی ROC[1]، در پایگاه اطلاعاتی اسکوپوس و مجلات حوزهٔ کتابداری و اطلاع‌رسانی داخلی نشان می‌دهد که در ایران معیارهای دقت و بازیافت کاربرد دارند و از لحاظ نظری معیار اِف، معیار ای[2] و به‌تازگی از طریق کُتُبی مانند *قلمروهایی نو در بازیابی اطلاعات* (بائزا-ییتس، و ریبرو- نتو ۱۳۸۵: ۱۲۸-۱۳۱) به متون راه یافته‌اند. در این مقاله سعی شده معیارها و مقیاس‌های ارزیابی ربط به تفکیک هر دو نوع نظام و با توجه به محیط پویای وب ارائه شود.

۲. سنجش اثربخشی یک نظام بازیابی اطلاعات

برای اندازه‌گیری اثربخشی نظام بازیابی اطلاعات به یک مجموعه اسناد مرکب از سه عنصر نیاز است (Manning, Raghavan and Schutze 2008):

۱. یک مجموعه از اسناد برای آزمون: برای ارزیابی اثربخشی هر نظام بازیابی اطلاعات، تعیین مجموعهٔ استاندارد اصول خود را دارد و تاریخ آن به اوایل دههٔ ۱۹۵۰ برمی‌گردد و در طول نزدیک به ۶۰ سال از زمان شروع آن، استفاده از مجموعه‌های آزمون[3] به استاندارد عملی ارزیابی بدل شده است (Sanderson 2010). از مشهورترین آنها می‌توان به مجموعه‌های استاندارد کرانفیلد[4]، ترک[5]، جی اُ وی دو[6]، ان‌تی‌سی‌آی‌ار[7]، کلف[8]، رویترز[9]، نیوزگروپ[10] اشاره کرد (Manning, Raghavan and Schutze 2008).

۲. دنباله‌ای از نیازهای اطلاعاتی[11] مانند سؤالات قابل بیان.

1. ROC: Reciever Operating Characteristic
2. E-measure
3. test collection
4. Cranfield
5. TREC
6. GOV_2
7. NTCIR
8. CLEF
9. Reuters
10. News Groups
11. A test suite of Information needs

۳. یک مجموعه از قضاوت‌های ربط: به‌صورت استاندارد یک ارزیابی دودویی، هم مربوط و هم غیرمربوط، برای هر دو جفت سند- سؤال (Manning, Raghavan and Schutze 2008): اگر میزان منابع مربوط به یک موضوع خاص در پایگاه مشخص باشد، ربط قابل سنجش است که چنین سنجشی فقط در پایگاه‌های اطلاعاتی پژوهش‌ساخته امکان‌پذیر است. به همین دلیل، تعیین مجموعهٔ استاندارد مانند ترِک اهمیت فوق‌العاده‌ای دارد.

۲-۱. معیارهای ارزیابی مجموعه‌های بازیابی رتبه‌بندی‌نشده[1]

مجموعه‌هایی که مدارک بازیابی‌شده را به‌صورت فهرستی از اقلام ارائه کرده و رتبه‌بندی نمی‌کنند.

۲-۱-۱. دقت و بازیافت: اغلب برای ارزیابی نظام‌های بازیابی اطلاعات به‌کار می‌روند (Raghavan, Bollman and Jung 1989). این دو مفهوم، مفاهیم پایه در بازیابی اطلاعات محسوب می‌شوند (Webber 2010) که در رشته‌ها و حوزه‌های گوناگون با اصطلاحات دیگری نیز نام‌گذاری می‌شوند. برای مثال، در پزشکی بازیافت را حساسیت (ROC Work 2011) و یا در داده‌کاوی، دقت را اطمینان[2] (Powers 2008) می‌نامند. البته مفهوم این اصطلاحات در این حوزه‌ها دقیقاً مشابه نیست، ولی می‌توان هم‌ارز دانست. برای تعریف دقیق دقت و بازیافت، دو نوع قضاوت وجود دارد: ۱- مدارک مربوط و ۲- مدارک نامربوط و در رابطه با این دو قضاوت مسئلهٔ دیگر این است: «ربط: سؤال در برابر نیاز اطلاعاتی[3] (Schutze 2013)». یعنی مربوط‌بودن و یا نبودن یک مدرک از چه نقطه نظری منظور است؟ ارتباط دقیق مدرک به سؤال یا به نیاز پژوهشگر؟ گاهی منبع بازیابی‌شده ۱۰۰ درصد به سؤال مربوط بوده و حاوی کلیدواژه‌های جستجو، هم در عنوان و هم در چکیده است، اما ربطی به نیاز اطلاعاتی کاربر ندارد. شوتز در این باره می‌گوید: رضایت کاربر تنها با این موضوع که نتایج بازیابی‌شده به نیاز او مربوط باشد نه به سؤال او، اندازه‌گیری می‌شود (Schutze 2013). با این حال، در بیشتر مواقع در ارزیابی ربط، باز هم قضاوت بر روی سؤال انجام می‌گیرد، نه نیاز اطلاعاتی و با دو معیار اصلی دقت و بازیافت

1. Evaluation of unranked retrieval sets
2. confidence
3. relevance: query vs. information need

سنجیده می‌شود.

فرمول ۱ و ۲. $\text{دقت} = \frac{\text{مدارک بازیابی‌شده مربوط}}{\text{مدارک بازیابی‌شده}}$ $\quad \text{بازیافت} = \frac{\text{مدارک مربوط بازیابی‌شده}}{\text{مدارک مربوط}}$

(webber 2010)

دقت و بازیافت با جزئیات بیشتری نیز قابل تعریف است:

جدول ۱. جدول انواع منابع بازیابی‌شده توسط نظام

	مربوط	نامربوط
بازیابی‌شده	مثبت واقعی *Tp*	مثبت کاذب *Fp*
بازیابی‌نشده	منفی کاذب *Fn*	منفی واقعی *tn*

۱. مثبت واقعی: اسناد مربوطی که بازیابی می‌شوند؛
۲. مثبت کاذب: اسناد نامربوط که مربوط بازیابی می‌شوند؛
۳. منفی کاذب: اسناد نامربوطی که بازیابی نمی‌شوند؛
۴. منفی واقعی: اسناد مربوطی که بازیابی نمی‌شوند.

بدین‌ترتیب دقت و بازیافت به شکل زیر تعریف می‌شوند (Evaluation in Information Retrieval 2007)

فرمول ۳ و ٤. $\textit{Recall} = \frac{tp}{tp+fn}$ $\quad \textit{Precision} = \frac{tp}{tp+fp}$

می‌توان دقت بسیار بالایی با بازیافت بسیار پایینی به‌دست آورد. این یک قانون مشخص در دقت و بازیافت است که افزایش یکی با کاهش دیگری همراه است. در این راستا معیار اندازه‌گیری دیگری وجود دارد که می‌توان با آن دقت و بازیافت را در مقابل

هم تعدیل (سبک و سنگین[1]) کرد (Schutze 2013). این معیار، معیار اِف[2] است.

۲-۱-۲. معیار اِف: این معیار ترکیبی از معیارهای دقت و بازیافت است (Kessler 2012) و در سال ۱۹۷۴ توسط ون ریجسبرگن[3] ارائه شده است و از معیارهای معمول ارزیابی، به‌خصوص در مواقع کار با مجموعه‌های غیرمتعادل است. این معیار، به‌ویژه زمانی که موارد مربوط در نظام بازیابی اطلاعات اندک هستند (Ye et al 2012) معمولاً به معیارهای دقت[4] ترجیح داده می‌شوند (Manning, Raghavan and Schutze 2008). این سه معیار، معیارهای معمولی در کارهای استخراج اطلاعات[5] همچون بازشناسی موجودیت مشخص[6] هستند (Tjong Kim Song and D Meulder 2003). معیار اِف میانگین توافقی دقت و بازیافت بوده (Sasaki 2007) و تعیین حداکثر مقدار برای اِف، تلاش به‌منظور یافتن بهترین ترکیب ممکن بین جامعیت و مانعیت می‌باشد (بائزا- ییتس، و ریبریو- نتو ۱۳۸۵: ۱۲۹).

$$F = \frac{2PR}{P+R}$$ **فرمول ۵. سنجش معیار اِف**

پاورز فرمول محاسبهٔ معیار اِف را به این شکل تعریف می کند: (Powers 2007)

$$F = \frac{1}{\alpha\frac{1}{P}+(1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P+R} \rightarrow \beta^2 = \frac{1-\alpha}{\alpha} \quad \beta \in [0,1] \text{ and thus } \beta^2 \in [0,\infty]$$ **فرمول ٦.**

اگر بخواهیم در آن معیار اِف متعادل[7] باشد (Powers 2007) باید $\beta = 1$ or $\alpha = 0$. که که در این صورت فرمول پیچیدهٔ بالا تبدیل به همان فرمول ساساکی (Sasaki 2007) می گردد. اگر $\beta<1$ تأکید بیشتر بر روی دقت نظام بازیابی است و اگر $\beta>1$ بازیافت نظام بیشتر مد نظر است، اندازهٔ هر سه معیار اندازه‌گیری دقت، بازیافت، و معیار اِف بین [۱، ۰] است، اما به‌طور معمول می‌توانند به‌صورت درصد هم بیان شوند (Manning, Raghavan and Schutze 2008).

1. trade off
2. F-measures
3. Van Rijsenbergen
4. accuracies
5. information extraction
6. named entity recognition
7. balanced F

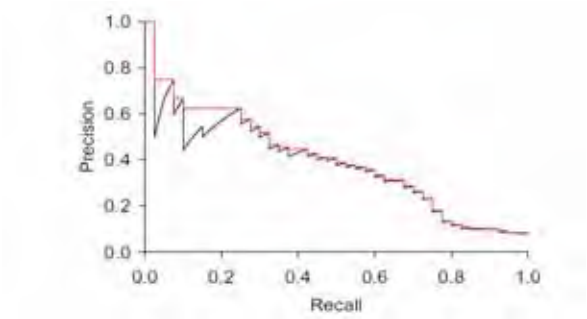
۲-۲. معیارهای ارزیابی نتایج بازیابی در نظام‌ها با نتایج رتبه‌بندی‌شده[1]

بسیاری از نظام‌های بازیابی، رکوردها و نتایج بازیابی اطلاعات را به‌صورت رتبه‌بندی‌شده ارائه می‌دهند. چنین نظام‌هایی، نیاز به سازوکاری دارند که از طریق آن، نزدیکی تطابق بین پرسش کاربر و یک سند را محاسبه نمایند. نتیجهٔ این محاسبه می‌تواند تعیین کند که اعضای یک مجموعه با چه ترتیبی به کاوش‌کننده نمایش داده شوند. این محاسبه، نظام را قادر می‌سازد که ربط یک سند را تخمین بزند و هدف این است که تخمین فوق همبستگی زیادی با قضاوت کاربر از ربط سند داشته باشد. نتیجهٔ این محاسبه، یعنی ارزش اختصاص‌یافته به نزدیکی تطابق بین پرسش و سند، ارزش وضعیت بازیابی[2] خوانده می‌شود. روش خاص ارزیابی ارزش وضعیت بازیابی، بستگی به مدل بازنمون اسناد مورد استفاده دارد (میدو، بویس[3]، کرافت[4] و باری[5] ۱۳۹۰: ۴۱۲) به تناسب مدل مورد استفاده در بازیابی اطلاعات، محاسبهٔ ارزش نتایج بازیابی رتبه‌بندی شده، می‌تواند دودویی (در مدل‌های منطق دو ارزشی مانند بولی) یا چندارزشی (در مدل‌های احتمالی و فازی) باشد. معیارهای کلاسیک ربط (دقت، بازیافت و اِف) معیارهایی بر پایهٔ مجموعه (مجموعه- پایه[6]) هستند. آنها با استفاده از مجموعه‌های غیرترتیبی از اسناد محاسبه می‌شوند، پس باید توسعه یابند، یا معیارهای جدیدی ایجاد شود تا بتوانند نتایج رتبه‌بندی‌شده را که در موتورهای جستجوی استاندارد هستند، ارزیابی کنند. در ادامه به این معیارها اشاره می‌شود.

۲-۲-۱. منحنی‌های بازیافت- دقت[7]: در زمینهٔ بازیابی رتبه‌بندی‌شده، مجموعه‌های مناسب اسناد بازیابی‌شده به‌طور طبیعی با k سند بازیابی‌شده بالایی[8] (k تعداد از اول فهرست) مشخص می‌شوند. در این مجموعه‌ها، ارزش دقت و بازیافت می‌تواند با نمایش یک منحنی بازیافت- دقت ترسیم شود (Manning, Raghavan and Schutze 2008).

---

1. evaluation of ranked retrieval results
2. Retrieval Status Value(RSV)
3. Boyce
4. Kraft
5. Barry
6. set-based measures
7. precision-recall curves
8. the top k retrieved documents



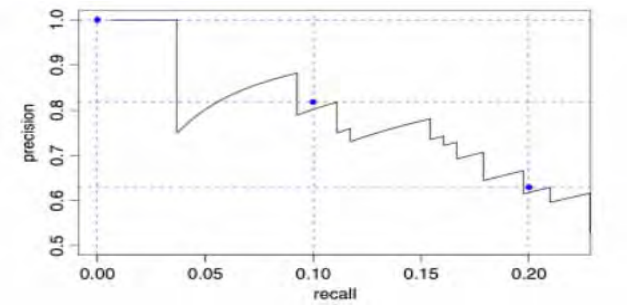
شکل ۱. بازیافت/ دقت در نظام‌ها با نتایج رتبه‌بندی‌شده (Manning, Raghavan and Schutze 2008)

۲-۲-۲. دقت افزوده (درونی)[1]: منحنی‌های دقت–بازیافت یک شکل دندانه‌دار مشخص دارند (شکل ۱). اگر سند بازیابی‌شده ($k+1$)[ام] مربوط نباشد، بازیافت همان بازیافت مربوط به سند k است، اما دقت کاهش پیدا می‌کند. اما اگر آن سند، مربوط باشد، هم بازیافت و هم دقت افزایش می‌یابد و بدین‌ترتیب تورفتگی‌های منحنی، بالا آمده و به طرف راست حرکت می‌کند. اگر این تورفتگی‌ها را برداریم، یک دقت افزوده محاسبه می‌شود. دقت درونی در یک سطح بازیافت r بالاترین دقت یافت شده در هر سطح r' است (Manning, Raghavan and Schutze 2008).

فرمول ۷. دقت افزوده $r' \geq r \rightarrow P_{interp}(r) = \max_{r' \geq r} p(r')$

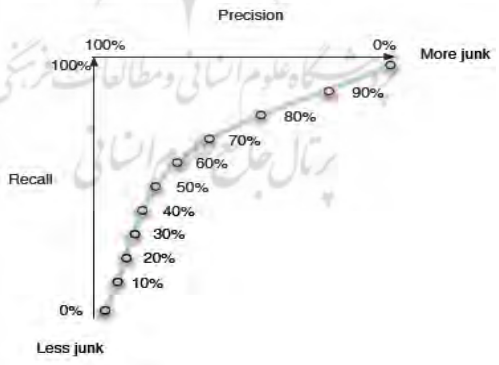
در برخی متون به‌جای درون‌یابی دقت[2]، هموارسازی[3] نیز به‌کار رفته است (Evaluation in Information Retrieval 2007). برای به‌دست آوردن این دقت در نقطه‌های بازیابی r ابتدا نقطهٔ مورد نظر را در محور x (خطوط نقطه‌چین عمودی) مشخص کرده و سپس بالاترین دقت (خطوط نقطه‌چین افقی) را در آن نقطه پیدا می‌کنیم. این نقطهٔ اوج، دقت درونی در نقطه r است که با نقطه‌های ثابت مشخص شده است (Lupo and et al. 2011). این نوع دقت‌ها، از نظر اطلاعات در موضوع یا نظام واحد بسیار غنی‌اند، اما برای مقایسه و استناد، اداره‌نشدنی هستند (Webber 2010).

1. interpolated precision
2. interpolation of precision
3. smoothing



شکل ۲. نحوهٔ محاسبهٔ دقت افزوده (Lupo et al. 2011)

۲-۲-۳. آزمون میانگین دقت افزوده (درونی) نقطهٔ ۱۱[1]: آزمون منحنی دقت- بازیافت حاوی اطلاعات بسیاری است، اما اغلب، این علاقه‌مندی وجود دارد که مقدار این اطلاعات به تعداد کمتری کاهش یابد، یا گاهی حتی به یک مورد برسد. راه سنتی برای انجام این کار، استفاده از میانگین دقت درونی نقطهٔ ۱۱ است. برای هر نیاز اطلاعاتی، دقت درونی در سطح بازیابی ۱۱، از صفر تا یک محاسبه می‌شود (Manning, Raghavan and Schutze 2008). می‌توان ارزیابی دقت را در نقطه‌های بازیافتِ مشخص انجام داد که میانگین دقت در نقطهٔ یازدهم بازیافت محاسبه می گردد. این نوع محاسبهٔ دقت بر پایهٔ اندازهٔ نوشته یا تعداد نتایج ارجاع داده‌شده و بازیابی‌شده نیست (Evaluation In IR 2008).



شکل ۲. آزمون میانگین دقت درونی نقطه ۱۱ (Evaluation In IR 2008)

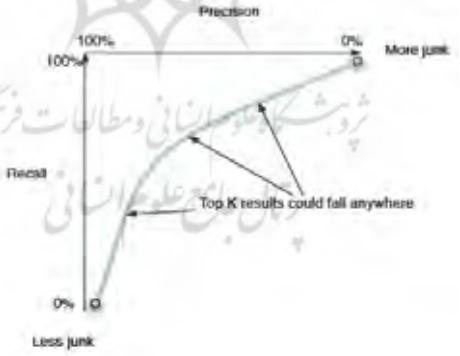
1. 11-point interpolated precision average

۲-۲-٤. میانگین دقت متوسط (مپ)[1]: دقت‌های محاسبه‌شده در بالا، همگی برای یک نیاز اطلاعاتیِ مشخص بودند. یعنی پرسشی به نظام داده شده و نتایجی ارائه گردیده است. با استفاده از فرمول محاسبهٔ دقت می‌توان دقت نظام را برای یک درخواست و نیاز اطلاعاتی محاسبه نمود. اما مسئله این است که برای ارزیابی نظام‌های بازیابی اطلاعات بر اساس مجموعه‌های آزمون استاندارد، تعداد زیادی درخواست گاهی تا ۱۵۰ نیاز اطلاعاتی به نظام ارائه می‌شود که هر کدام، دقت خود را دارد (Manning, Raghavan and Schutze 2008) و برای هر سؤال، می‌توان معیاری به نام میانگین دقت[2] محاسبه کرد (Sujatha and Dhavachelvan 2011) که تعدد دقت‌ها می‌تواند کار ارزیابی نظام را با دشورای روبه‌رو کند. در اینجا از میانگین دقت متوسط استفاده می‌شود. این نوع دقت، استانداردترین نوع محاسبهٔ دقت در مجموعهٔ آزمون ترِک[3] است که یک نمره واحد از کیفیت در طول سطوح بازیافت نظام ارائه می‌کند و ثبات بالایی دارد (Manning, Raghavan and Schutze 2008). این نوع دقت، تقریباً برابر با سطح زیر منحنی است. میانگین دقت متوسط به‌صورت یک معیار اثربخشی معتبر پذیرفته و استفاده شده است (Cormack and Lynam 2006).

شکل ۳. دقت میانگین متوسط (Evaluation In IR 2008)

1. mean average precision
2. average precision
3. TREC:Text REtrieval Conference

۲-۲-۵. دقت در K[1]: اگر کاربران و یا قضاوت‌هایی وجود داشته باشد که با مربوط‌بودن یا نبودن مدارک بازیابی‌شده موافقت کنند، محاسبهٔ دقت کار آسانی است؛ البته با این فرض که اندازهٔ مجموعهٔ بازیابی‌شده قابل کنترل باشد. در نظام‌های بازیابی اطلاعات که اسناد را به‌صورت رتبه‌بندی‌شده ارائه می‌دهند، می‌توان اسناد بازیابی‌شده را با قراردادن یک آستانهٔ بالاتر (برای مثال جستجو در ۱۰۰ مورد اول و یا ۲۰ مورد بالای لیست) کاهش داد. این همان دقت در **K** است (Sujatha and Dhavachelvan 2011). فاکتورهای قبلی، دقت را در همهٔ سطوح بازیابی محاسبه می‌کند ولی در کاربردهای ویژه در موتورهای جستجو، این نوع محاسبهٔ دقت، به کاربر زیاد مربوط نیست. این موضوع که اسناد مربوط در صفحهٔ اول تا ۳ صفحهٔ اول جستجو وجود دارد، همیشه برای کاربر وجود داشته (Manning, Raghavan and Schutze 2008) و همین عامل محاسبهٔ دقت در نقطهٔ K است. چون ثابت شده که کاربران، بیشتر از یک یا دو صفحهٔ اول فهرست نتایج بازیابی را بررسی نمی‌کنند (Evaluation In IR 2008)، پس محاسبهٔ دقت، بررسی تعداد محدودی انتخاب از فهرست نتایج بازیابی شده است. این نوع محاسبه، اگر چه دارای ثبات کمتری نسبت به فاکتورهای قبلی است، اما نیازی به تخمین اندازهٔ اسناد مربوط در پایگاه‌های اطلاعاتی بزرگ در هنگام ارزیابی ندارد.

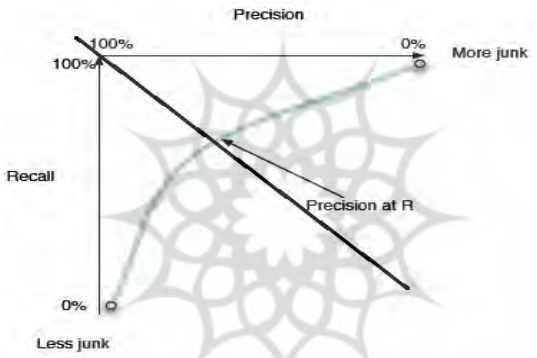


شکل ٤. محاسبهٔ دقت در K (Evaluation In IR 2008)

1. precision at K

۲-۲-۶. دقت R'[1]: یک گزینه که مشکل مورد قبلی، یعنی ثبات پایین دقت را کم می‌کند، استفاده از دقت R است. در این دقت لازم است که مجموعه‌ای از اسناد مربوطِ شناخته‌شده را داشته باشیم که با آن دقت R مورد اول فهرست بازیابی‌شده را به‌دست آوریم (Manning, Raghavan and Schutze 2008). این نوع دقت، دقت نظام در نقطه R'[2] در رتبه‌بندی نتایج برای یک سؤال است که برای آن در نظام، R سند مربوط وجود دارد (Wikipedia 2011). به‌عبارت دیگر، اگر r سند مربوط در میان R سند بازیابی‌شده بالایی(اول) فهرست وجود دارد، محاسبهٔ دقت R این‌گونه است:

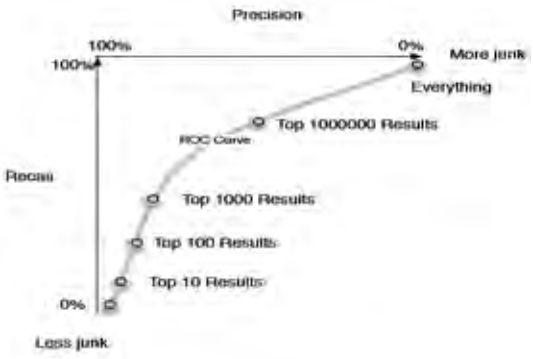
فرمول ۸. $\text{R}-\textit{Precision} = \frac{r}{R}$ **(Craswell 2013)**



شکل ۵. محاسبهٔ دقت در R (Evaluation In IR 2008)

۲-۲-۷. منحنی ROC و حساسیت[2]: معیار دیگر، منحنی ROC یا همان مشخصهٔ عامل گیرنده است که البته شناختن آن به بسیاری از افراد کمکی نمی‌کند. این منحنی نرخ مثبت درست[3] را در مقابل نرخ مثبت اشتباه[4] ترسیم می‌کند (Manning, Raghavan and Schutze 2008).

1. R-precision
2. Sensitivity
3. True-positive (TP)
4. false-positive (FP)



شکل ٦. محاسبهٔ دقت در ROC (Evaluation In IR 2008)

در علوم پزشکی، ویژگی عامل گیرنده یا همان ROC از پردازش سیگنال‌ها به امانت گرفته شده تا تبدیل به استانداردی برای ارزیابی و مقایسهٔ نرخ مثبت درست و نرخ مثبت اشتباه شود. در علوم رفتاری معمولاً از اختصاصیت[1] (خاص‌بودن) و حساسیت[2] استفاده می‌شود. فنون جایگزین مانند دقت مرزی[3] دارای مزیت‌هایی است، اما هنوز هم در آن سوگیری‌هایی وجود دارد (Powers 2008). در هر نظام بازیابی اطلاعات، دقت و بازیافت دو عامل ارزیابی اصلی محسوب می‌شوند، ولی آیا آنها معیارهای کامل ارزیابی نظام‌اند؟ آن هم برای همه نظام‌های بازیابی؟ به‌طور مثال، چقدر برای ارزیابی تست‌های ارزش تشخیصی در پزشکی می‌توان فقط به دقت و بازیافت تکیه کرد. آیا معیارهایی دقیق‌تر وجود دارد؟ در حوزهٔ بازیابی به‌خصوص پزشکی ۴ اصطلاح وجود دارد. اصطلاحات زیر با یک مثال در حوزهٔ پزشکی توضیح داده‌می‌شود:

۱. مثبت واقعی یا همان حساسیت و یا بازیافت: به معنای بیمارانی هستند که واقعاً بیمار تشخیص داده می‌شوند؛

۲. منفی واقعی یا همان خاص‌بودن (اختصاصیت): افراد سالمی که سالم تشخیص داده می‌شوند؛

---

۱. این اصطلاح رایجی میان متخصصان حوزهٔ پزشکی است و از ادبیات آنها گرفته شده است.

2. specificity & sensitivity

3. rand accuracy

۳. منفی کاذب[1]: بیمارانی هستند که سالم تشخیص داده می‌شوند؛
۴. مثبت کاذب[2]: افراد سالمی که بیمار تشخیص داده می‌شوند (Metz 2006).

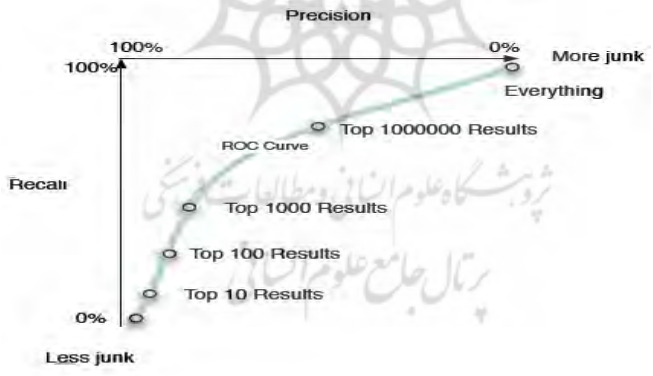
برای این ۴ مفهوم می‌توان ۴ کسر یا تابع تعریف کرد (ROC Work 2011):

(1)$TPF^{3} \equiv TP/TDP^{4} \rightarrow (0 \leq TPF \leq 1)$ (2) $TNF \equiv TN/TND5 \rightarrow (0 \leq TNF \leq 1)$

(3) $FPF \equiv FP/TND \rightarrow (0 \leq FPF \leq 1)$ (4) $FNF \equiv FN/TDP \rightarrow (0 \leq FNF \leq 1)$

فرمول ۹. TNF=1-FPF and FNF=1-TPF

این چهار مفهوم، منحنی ROC را می‌سازند. اگر حوزهٔ زیر نمودار منحنی برابر با یک باشد، به معنی یک پیش‌بینی کامل است. یعنی همهٔ موارد مثبت در بالای همهٔ موارد منفی قرار می‌گیرد. اگر سطح زیر نمودار برابر با ۵ باشد، یک پیش‌گویی تصادفی از دقت است که در آن هیچ ارتباطی میان ارزش‌های پیش‌گویی‌شده و درست وجود ندارد (Geng, Wong and Li 2004). یک انتخاب برای ارزیابی دقت آن است که میانگین نمرات دقت در هر نقطه در منحنی، جداگانه محاسبه شود. اما، مسئلهٔ اساسی این است که کدام نقطه؟ (Evaluation In IR 2008).



نمودار ٦. محاسبهٔ دقت در ROC (Evaluation In IR 2008)

1. false-negative (FN)
2. false-positive (FP)
3. fraction
4. total number of diseased patient
5. total number of none-diseased patient

جدول ۲. دسته‌بندی معیارهای ارزیابی نظام‌های بازیابی اطلاعات

معیارهای ارزیابی نظام‌های رتبه‌بندی‌نشده	معیارهای ارزیابی نظام‌های رتبه‌بندی‌شده
دقت	منحنی‌های دقت و بازیافت
بازیافت	دقت افزوده و میانگین دقت افزوده در نقطهٔ ۱۱
معیار اِف	میانگین دقت متوسط
صحت	دقت در K و دقت در R
-------	منحنی ROC
-------	افزایش تجمعی یا افزایش تجمعی نزولی نرمال

۳. مقیاس اندازه‌گیری

بحث در مورد معیارهای ارزیابی ربط، بسیار طولانی است، اما همیشه مسئلهٔ مهم در ارزیابی علاوه بر معیار ارزیابی، مقیاس‌ها هستند. استفاده از مقیاس‌های درست و دقیق در ارزیابی می‌تواند به‌خوبی راهگشا باشد. در بسیاری از پژوهش‌ها از مقیاس‌های دودویی استفاده شده است. برای مثال، ترِک که معروف‌ترین مرجع برای تشویق پژوهش در بازیابی اطلاعات در مجموعه‌های متنی بزرگ است (NIST 2013).

همیشه مقیاس دودویی را به‌کار می‌برد، به معنای آنکه قضاوت‌کنندگانِ ربط یا مدرکی را مرتبط می‌دانند یا نمی‌دانند (Voorhees & Harman 2001 نقل در Vaughan 2004 679) (نقل در خالویی ۱۳۸۷: ۳۷) و (Cooper & Chen 2001) همچنین (Hawking, Griffiths 2001 Craswell, Baily &) مقیاس دودویی مرتبط و نامرتبط را برگزیده‌اند.

اما همان‌طور که گفته شد ربط، نسبی و پویاست و پویابودن ربط به معنی آن است که استنباط‌های به‌عمل آمده از ربطِ یک مدرک اطلاعاتی برای یک کاربر در طول زمان دستخوش تغییر می‌گردد (Borlund 2003 نقل در داورپناه ۱۳۸۶: ۲۰۷). بدین‌ترتیب نظریهٔ دوبخشی ربط از طرف پژوهشگرانی مانند ربوتسون، ساراسویک و جینز مورد انتقادهای فراوانی واقع شد (Yao 1993: 134). این موضوع باعث شد که کم‌کم قضاوت‌های چندسطحی در مقابل قضاوت‌های دو ارزشی خود را نشان دهند. چون قضاوت‌های ربط دو ارزشی هیچ اهمیتی به ربط‌های جزئی و تا حدی مربوط به منابع بازیابی‌شده نمی‌دهد،

چو[1] و روزنتال[2] (۱۹۹۶) از درجهٔ ربط ۳ سطحی (مربوط، تا اندازه‌ای مربوط، و نامربوط) بهره بردند (Chu & Rosental 1996) در همان سال دینگ[3] و مارچیونینی[4] یک مقیاس ۶ درجه‌ای را در پژوهش خود پیشنهاد داده و حتی لینک به دیگر منابع را نیز، که موضوع بسیار مهمی در پایگاه‌های اطلاعاتی بود، در نظر گرفتند. آنها سه نوع دقت متفاوت با استفاده از این مقیاس‌ها تعریف کردند. سو، چن و دانگ مقیاس ۵ درجه‌ای را تعریف کردند و همبستگی رتبه‌بندی را با رتبه‌بندی‌های موتورهای جستجو سنجیدند (Su, Chen & Dong 1998). سال بعد گوردون و پاتاک درجهٔ سه سطحی چو و روزنتال را به ۴ سطحی افزایش دادند تا دقت ارزیابی بالاتر رود (Gordon & Pathak 1999). این ۴ درجه بسیار مربوط، تا اندازه‌ای مربوط، تا اندازه‌ای نامربوط، بسیار نامربوط بود (Vaughan 2004).

از این نوع مقیاس‌ها می‌توان به مقیاس سه درجه‌ای ساراسویک (۱۹۶۹) و جینز (۱۹۹۱)، ۹ درجه‌ای کترا و کدر (۱۹۶۷)، ۱۱ درجه‌ای ریز و شولتز (۱۹۹۷)، ۱۳ درجه‌ای هوارد (۱۹۹۴) اشاره کرد (اخوتی ۱۳۸۷، ۳۹)[5]. اما با وجود چند درجه‌ای‌شدن مقیاس‌ها و دقیق‌ترشدن معیارهای ارزیابی نظام‌های بازیابی، مسائل دیگری هم مطرح است. پیشرفت‌ها در حوزهٔ فناوری، نظام‌های بازیابی اطلاعات را تبدیل به زنجیره‌های دانشی کرده که به مدد فرامتن و لینک‌های ورودی و خروجی، شبکه‌ای از گره‌های علمی را به‌وجود آورده است. ارزیابی نظام کار ساده‌ای نیست. در پایگاه اطلاعاتی پابمد[6] در هر بازیابی اطلاعات، اطلاعات، هر مورد بازیابی‌شده به دو نوع اطلاعات دیگر پیوند دارد. فهرست منابع مقاله که به‌صورت فرامتن (در بیشتر موارد) در انتهای مقالات تمام‌متن وجود دارند و فهرست مقالات مرتبط که کاربران به‌راحتی می‌توانند از آنها نیز استفاده نمایند. در چنین شرایطی آیا یک مقالهٔ بازیابی‌شده در ابتدای فهرست رتبه‌بندی نظام پاب‌مد، که به موضوع مورد جستجوی کاربر مرتبط است، نسبت به یک مقاله در رتبهٔ ۱۸ فهرست رتبه‌بندی، که ارتباط جزئی با یک مقاله دارد، اما به یک مقالهٔ کاملاً مرتبط با موضوع به‌صورت فرامتن پیوند

---

1. Chu
2. Rosental
3. Ding
4. Marchionini

۵. متأسفانه در مقالهٔ اخوتی رفرنس‌های مربوط به این مقالات ذکر نشده بود و تلاش پژوهشگران نیز به‌دلیل نبودِ اطلاعات کافی به نتیجه‌ای نرسید.

6. PubMed

خورده است، ارجحیت دارد یا نه؟ آیا این دو، ارزش مساوی دارند و یا ارزش مقالهٔ دوم نسبت به مقالهٔ اول بیشتر است؟ هر چند که این موضوع به قضاوت کاربر بستگی دارد، اما گاهی کاربران به‌دلیل مشغله و یا به‌علت کمبود وقت و یا ناآگاهی از وجود چنین امکاناتی ترجیح می‌دهند تنها فهرست بازیابی‌شده را قضاوت کنند و گاهی موضوع پژوهش آنقدر تخصصی و حیاتی است که کاربران زمان بیشتری برای جستجو اختصاص داده و لینک‌ها و فرامتن‌ها را نیز مورد بازبینی قرار می‌دهند. در این شرایط برای سنجش، باید معیارهای ارزیابی منتخب و ویژگی‌های فرامتنی را نیز در نظر گیرند. هر چند که ارزیابی در این حالت، سخت و مشکل می‌شود، اما دقیق‌تر خواهد بود. در این خصوص گوئیدزا و چیگنل، ۴ مقیاس (کاملاً مربوط، تا اندازه‌ای مربوط، تا حد کمی مربوط و نامربوط) را معرفی کرده و به این مقیاس‌ها نمره‌های ۳، ۲، ۱ و صفر دادند و بر اساس آنها ۴ نوع دقت کامل[1]، دقت برتر[2]، دقت مفید[3] و دقت واقعی[4] را اندازه گیری نمودند (Gwizdka and Chignell 1999).

$$precFull(minFnHits) = \frac{\sum_{i=1}^{minFnHits} score_i}{minFnHits * maxHitScore}$$

فرمول ۱۰. محاسبهٔ دقت کامل

در فرمول دقت کامل، فرض بر این است که نمره‌های ربط قابل جمع هستند. اما، هنوز تحلیلی نظری انجام نشده تا چنین ادعایی را اثبات کند. در این نوع دقت باید فرض کرد که دو سند با نمرهٔ ۳ ارزش برابری با ۶ سند با ارزش ۱ دارند که چنین ادعایی در حد نظریه می‌باشد. این در حالی است که دقت‌های دیگر قابل اندازه‌گیری است:

$$precBest(minFnHits) = \frac{count\_of(score_i = 3)}{minFnHits}$$

فرمول ۱۱. نحوهٔ محاسبهٔ دقت برتر

1. Full Precision
2. Best Precision
3. Useful Precision
4. Objective Precision

فرمول ۱۲. نحوهٔ محاسبهٔ دقت مفید

$$precUse(minFnHits) = \frac{\sum_{i=1}^{minFnHits} count\_of(score_i \geq 2)}{minFnHits}$$

فرمول ۱۳. نحوهٔ محاسبهٔ دقت واقعی

$$precObj(minFnHits) = \frac{\sum_{i=1}^{minFnHits} count\_of(score_i > 0)}{minFnHits}$$

در بررسی‌ها مشخص شد که این فرمول‌ها، یکی از معدود فرمول‌های اندازه‌گیری دقت نظام‌های بازیابی اطلاعات بر اساس فرامتن و لینک‌های موجود در متن است. با استفاده از این دقت‌ها، ۴ نوع دقت دیگر نیز معرفی شد که به دقت متغیر یا دقت تفاضلی[1] معروف است، چون بیشتر کاربران تمایل دارند از میان انبوه مدارک بازیابی‌شده، ۱۰ تا ۲۰ نتیجهٔ اولیه را بازبینی کنند. پس، منطقی است که محاسبات دقت به ۲۰ یافتهٔ اول محدود شود. با این توضیح، برای ارزیابی دقت تفاضلی یا دقت متغیر می‌توان از فرمول زیر استفاده نمود.

فرمول ۱٤.

$$dpObj(1, minF20Hits) = precObj(1, minF10Hits) - precObj(minF10Hits, minF20Hits)$$

اگر عدد به‌دست آمده از صفر بزرگتر باشد، به معنای آن است که تعداد مدارک مربوط در ۱۰ مورد اول بازیابی بیشتر از ۱۰ مورد دوم است؛ و اگر عدد به‌دست آمده برابر با صفر باشد، به معنای آن است که تعداد مدارک مربوط در ۱۰ مورد اول بازیابی برابر با ۱۰ مورد دوم است؛ و اگر عدد به‌دست آمده از صفر کوچکتر باشد، به معنای آن است که تعداد مدارک مربوط در ۱۰ مورد اول بازیابی کمتر از ۱۰ مورد دوم است. دقت متغیر را می‌توان برای دقت‌های برتر و مفید نیز محاسبه نمود و در فرمول مربوطه از دقت‌های مفید و برتر به‌جای دقت واقعی استفاده کرد. چنین ارزیابی‌ای بسیار سخت و مشکل به نظر می‌رسد، چون تمام کاربران زمان کافی برای بازدید از نتایج بازیابی‌شده به‌همراه لینک‌ها و فرامتن‌ها را ندارند، اما این موضوع، که یک پایگاه اطلاعاتی نیز بتواند چنین امکانی، یعنی

[1]. Differential Diagnosis

ایجاد لینک به دیگر منابع یا مقالات مربوط یا فهرست منابع یک مقاله را فراهم کند، بسیار موضوع مهمی است. بنابراین، ارزیابی نظام باید با در نظر گرفتن همهٔ جوانب و همه جنبه‌های موجود انجام شود.

٤. معیارهای ارزیابی ربط در پژوهش‌های داخلی

یکی از عمده پژوهش‌های انجام‌شده در حوزهٔ علم اطلاعات و دانش‌شناسی(کتابداری و اطلاع‌رسانی) ارزیابی میزان دقت و بازیافت پایگاه‌های اطلاعاتی است. با استفاده از کلیدواژه‌های «ربط»، «ارزیابی بازیابی»، «بازیابی اطلاعات»، «دقت»، «بازیافت»، «معیار اِف» و «منحنی ROC» بر روی مجلات کتابداری و اطلاع‌رسانی بررسی دقیقی انجام گرفت. یافته‌ها نشان داد که فقط معیار دقت، بازیافت و میانگین دقت مورد بررسی قرار گرفته است. بررسی پژوهش‌های ایرانی موجود در اسکوپوس نیز این موضوع را تأیید کرد. شاید دلیل این امر، ضعف مبانی نظری در این حوزه باشد. همین بررسی نشان داد که از جنبهٔ نظری نیز در ارتباط با مقیاس‌های ربط فقط در پژوهشی از خالویی (۱۳۸۷) و کتاب نظریه‌های رفتار اطلاعاتی از داورپناه (۱۳۸۷) به انواع مقیاس‌ها اشاره شده است. معیار اِف، معیار ای، تازگی، جامعیت نسبی و دستاورد جامعیت از طریق کتاب *قلمروهای نو در بازیابی اطلاعات* به متون فارسی راه یافته‌اند. اما از آنجا که استفاده از پایگاه‌های اطلاعاتی وبی در طی سال‌های اخیر در ایران رشد چشمگیری یافته و هزینه‌های زیادی بر کتابخانه‌ها و مراکز اطلاع‌رسانی تحمیل می‌کند، جا دارد این پایگاه‌ها با معیارهای دقیق ارزیابی، سنجش شوند تا هم بتوان هزینه‌ها را توجیه نمود و هم کاربران را به اطلاعات مناسب و به‌هنگام رهنمون کرد.

٥. بحث و نتیجه‌گیری

انواع مختلفی از پایگاه‌های اطلاعاتی در در حوزهٔ وب وجود دارند که به‌دلیل طبیعت پویای وب به‌سرعت اعجاب‌آوری در حال پیشرفت و توسعه کمی و کیفی هستند. انتخاب از میان انبوه این پایگاه‌ها مشکل، اما اجتناب‌ناپذیر است. بنابراین، باید با استفاده از معیارها و مقیاس‌های ارزیابی مناسب به انتخاب درست آنها پرداخت تا بتوان از یک جهت، نیروی انسانی و منابع مالی را به‌درستی مدیریت نمود و از طرف دیگر، رضایت

کاربران را با مناسب‌ترین منابع به‌دست آورد. بسیاری از ارزیابی‌های نادرست در این حوزه بیش از آنکه به نبودِ ابزار اندازه گیری مربوط باشد، ناشی از ناآگاهی و عدم شناخت ابزارهای مناسب است. به‌کارگیری معیار و مقیاس درست ارزیابی باعث می‌شود که نتایج معتبر و قابل اطمینان به‌دست آید و بتوان از آن در تصمیم‌گیری‌ها با اطمینان بهره گرفت و از آنها دفاع کرد. از طرف دیگر، طراحی هر نظام اطلاعاتی مستلزم شناخت نقاط قوت و ضعف نظام‌های قبلی است که این نقاط قوت وضعف در ارزیابی‌های درست مشخص می‌شود. در این مقاله سعی شد تا با نگاهی جامع به متون موجود در حوزهٔ معیارهای ارزیابی نظام‌های بازیابی اطلاعات با تأکید برآنچه که دربارهٔ آنها می‌دانیم، آنچه را که تاکنون معرفی نشده و یا مورد غفلت قرار گرفته با تکیه بر نظام‌های تحت وب به جامعه پژوهشگران معرفی نماییم.

٦. فهرست منابع

اخوتی، مریم. ۱۳۸۳. مفهوم ربط در نظام بازیابی اطلاعات. *اطلاع‌شناسی* ۲ (۱): ۲۳-۴۵.

بائزا – ییتس، ریکاردو، برتیه ریبرو– نتو. ۱۳۸۵. *قلمروهای نو در بازیابی اطلاعات*. ترجمهٔ علی‌حسین قاسمی. تهران: چاپار؛ دبیزش.

حسن‌زاده، محمد. ۱۳۸۳. تأثیر مدل‌های بازیابی اطلاعات بر میزان ربط. *اطلاع‌شناسی* ۲ (۱): ۶۳-۸۹.

خالوئی، مرضیه. ۱۳۸۷. ربط و مفهوم آن در بازیابی اطلاعات. *علوم و فناوری اطلاعات* ۲۳ (۳): ۱۰۵-۱۱۸.

داورپناه، محمدرضا. ۱۳۸۶. *ارتباط علمی: نیاز اطلاعاتی و رفتار اطلاع‌یابی*. دبیزش: چاپار.

میدو، چارلز تی، برت‌ارت بویس، اچ کرافت و کارول باری. ۲۰۰۷. *نظام‌های بازیابی اطلاعات متنی*. ترجمهٔ نجلا حریری. ۱۳۹۰. تهران: چاپار.

Bouramoul, A, M. Kh Kholadi and B-L. Doan. 2011. Using Context to Improve the Evaluation of Information Retrieval System. *International Journal of Database Management Systems* 3 (2):22-39.

Borlund, P. 2003. The concept of relevance in IR. *Journal of the American Society for Information Science and Tchnology*, 54 (10): 913-925.

Carterette, B. A. 2008. Low-Cost and Rubust Evaluation of Information retrieval Systems. Ph.D Thesis, University of Massachusetts. http://ir.cis.udel.edu/~carteret/papers/thesis.pdf (accessed Feb. 24, 2013).

Chu, H & M Rosental. 1996. *Search Engines for the Word Wide Web: A Comprative Study and Evaluation Methodology*. Proceedings of the 57th Annual Meeting of the American Society for Information Science. *Journal Today* .127-136.

Cooper, M and H Chen. 2001. Predicting the Relevance of a Library Catalog Search. *Journal of*

the American Society fo Information Science and Technology 52 (10):813-827.

Cormack, T. T and Thomas R Lynam. 2006. *Statistical Precision of Information Retrieval Evaluation.* Paper presented At *SIGIR*, USA, Washington, ACM:533-540 http://itls.usu.edu/~beijxu/articles/statisticalPrecision.pdf (accessed Sep. 24, 2013).

Craswell, N, Ling Liu, and Özsu Tamer. 2013. *R-Precision* .SpringerReference. http://www.springerreference.com/docs/html/chapterdbid/63597.html (accessed Sep. 24, 2013).

Ding, W. and G. Marchionini. 1996. *A Comprative Study Search Engine Service Performance.* Propceeding of the 59th annual of Meeting American Society for Information Science. Medford: *Journal Today*.136-142

Evaluation in Information Retrieval.2007. Paper presented At Seminar für Sprachwissenschaft. International Studies in Computational Linguistics. http://www.sfs.uni-tuebingen.de/~parmenti/slides/slides9-1x4.pdf (accessed May. 20, 2013).

Evaluation In IR.. 2008. *Donald Bern School of Information and Computer Science* http://www.ics.uci.edu/~djp3/classes/2008_01_01_INF141/Lectures/Slides26.pdf (accessed Feb. 3, 2013).

Geng, Guang Gang, CHun-Heng Wang and Qui-Dan Li. 2004. *Improving of Spamdexing Detection via a Tow-Stage Classification Strategy*. Proceedings of the 4th Asia information retrieval conference on Information retrieval technology .Springer-Verlag Berlin. 356-364. http://link.springer.com/chapter/10.1007%2F978-3-540-68636-1_34#page-1 (accessed Feb. 6, 2013).

Gray, Alexander. 2011. *Web Search and text Mining: Evaluation in Retrieval Information.*Georgia Institute of Technology, College of Computing. http://www.cc.gatech.edu/~agray/6240spr11/08eval.pdf (accessed Sep. 21, 2013).

Gwizdka, Jacekand and Mark Chignell. 1999. *Towards Information Retrieval Measures for Evaluation of Web Search Engines.* http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.15.3212&rep=rep1&type=pdf (accessed Feb. 26, 2011).

Hagglund, Ulrika. 2004. *Modelling Information Objects to Promote Universal Matching* . Sweden:Umea. www.cs.umu.se/research/reports/2004/006/part1.pdf (accessed May 6, 2013).

Hawking, D, N Craswell, P Bailey and K Griffiths. 2001. Measuring Search Engine Quality *Information Retrieval* 4 (1): 33-59.

Kessler, Wiltrud. 2012. *Evaluation of Text Classification*. http://www.ims.uni-stuttgart.de/institut/mitarbeiter/kesslewd/lehre/sentimentanalysis12s/ml_evaluation.pdf (accessed May 6, 2013).

Lupo, M., K. Mayer, J. Taitand & A. J. Trippe. 2011. *Current Challenges in Patent Information Retrieval* The Information Retrieval Series Springer.

Manning, Ch, P. Raghavan and H. Schütze. 2008. *Introduction to Information Retrieval, Cahpter 1: Evaluation In Information Retrieval.* Cambridge: Cambridge University Press. http://nlp.stanford.edu/IR-book/pdf/08eval.pdf (accessed May 9, 2013). 152-162

Meadow, C. 1992. *Text Information Retrieval System.* Orlando: Academic Press, Inc.

Metz, C. E. 2006. Reciever Operating Charachteristic analysis: A Tool for Quantitative Evaluation of Observer Performance and Imaging System. *Journal of the American College of Radiology*: 3 (6): 413-422

http://www.uphs.upenn.edu/radiology/education/resources/documents/Receiver-operating-characteristic-analysis-tool.pdf (accessed May 9, 2013).

Moen, Pirjo. 2010. Information Retrieval Methods: Evaluation Information Retrieval.

NIST. 2013. TREC: Text REtrieval Conference http://trec.nist.gov/. (accessed May 9, 2013).

Powers, D. M. W. 2007. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies* 2 (1) 37–63. (accessed May 9, 2013).

_____ 2008. Evaluation Evaluation: A Monte Carlo Study. http://david.wardpowers.info/BM/ECAIacc-Evaluation_Evaluation.pdf (accessed May 9, 2013). http://www.bioinfo.in/uploadfiles/13031311552_1_1_JMLT.pdf

Raghavan, Vijay, Peter Bollmann, and Gwang Jung S.1989. A critical investigation of recall and precision as measures of retrieval system performance. *ACM Transactions on Information Systems (TOIS)*. 7 (30), 205-209 http://dl.acm.org/citation.cfm?id=65945 (accessed in Jan. 21, 2014)

ROC Task. 2011. http://www.devchakraborty.com/Receiver%20operating%20characteristic.pdf. (accessed May 9, 2013).

Sanderson, M. 2010. Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends &in Information Retrieva*l 4 (4): 247-375 http://www.nowpublishers.com/articles/foundations-and-trends-in-information-retrieval/INR-009 (accessed May 22, 2013).

Saracevic, T. 1995. *Evaluation of Evaluation in Information Retrieval* .Proceedings of the Eighteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Assiciation for Computing Machinary. 136-148 http://comminfo.rutgers.edu/~muresan/IR/Docs/Articles/sigirSaracevic1995.pdf (accessed Aug 21, 2013).

Sasaki, Y. 2007. *The Truth of the F-measure*. http://www.cs.odu.edu/~mukka/cs795sum09dm/Lecturenotes/Day3/F-measure-YS-26Oct07.pdf

Schamber, L., M. B Eisenberg and M. S. Nilan. 1990. A re-examination relevance: Toward a dynamic, situational definition. *Information Processing & Management* 26: 755–77 http://www.sciencedirect.com/science/article/pii/030645739090050C (accessed Jan. 21, 2014)

Schutze, H. 2013. Introduction to Inforation Retrieval: Relevance Feedback and Quary Extention. http://www.cis.uni-muenchen.de/~hs/teach/13s/ir/pdf/09expand.pdf (accessed Aug. 21, 2013).

Snášel, V., A. Abraham, S. Owais, J. Platoš and P. Krömer. 2009. *Optimizing Information Retrieval Using Evolutionary Algorithms and Fuzzy Inference System* 204.299-324 http://link.springer.com/content/pdf/10.1007%2F978-3-642-01088-0_13.pdf (accessed May 9, 2013).

Su, L.T., H Chen and X Dong. 1998. *Evaluation of Web-Based Search engines From the End-Users Perspective: A Pilot Study*. Prceeding of 16th Annual Meeting of the American Society for Information Science 348-364.

Sujatha, Pothula Dhavachelvan. 2011. Precision at K in Multilingual Information Retrieval. *International Journal of Computer Application 24 (9):40-43*. http://www.ijcaonline.org/volume24/number9/pxc3873929.pdf (accessed Apr. 1, 2012).

Tjong Kim Song, E.F and F D Meulder. 2003. *Introduction to the CoNll-2003 shared task: language -independent entity recognition*. HLT-NAACL:North American Chapter of the Association for Computational Linguistics. Edmonton http://mirror.aclweb.org//hlt-naacl03/call.html (accessed 20 Aug. 2013).

Van Rijsenbergen, C. J. 1974. Fundation of Evaluation. *Journal of Documentation 3(4): 365-373*

Vaughan, L. 2004. New Measurement For Search Engine Evaluation Proposed and Tested. *Information Processing and Management* 40: 677-691. http://www.jasonmorrison.net/iakm/cited/Vaughan_new_measurements.pdf (accessed Feb. 21, 2013).

Voorhiees, E. M and Harman Donna. 2001. *Overwiev Of TREC 2001*. Paper Presented At 10th Text Retrieval Conference(TREC 2001). Gaithersburg, Maryland *NIST Special Publication* 15 (1): 500-250 http://trec.nist.gov/pubs/trec10/papers/overview_10.pdf (accessed Feb. 23, 2013).

Webber, W. E. 2010. Measurment in Information Retrieval Evaluation. Ph.D Thesis. The university of Melburne http://www.umiacs.umd.edu/~wew/wew-thesis-PhD.pdf (accessed Sep. 24, 2013).

Wikipedia. 2011. IR Evaluation. http://en.wikipedia.org/wiki/IR_evaluation (accessed Apr. 27, 2013).

Yao, Y.Y. 1995. Measuring retrieval Effectiveness Base on User Preference of Documents. Issue. *Journal of the American Society for Information Science* 46 (2): 133-145. http://www2.cs.uregina.ca/~yyao/PAPERS/jasis_ndpm.pdf (accessed Jan 22, 2014)

Ye, Nan, K. M. A. Chai, Wee Sun Lee and Hai Leong Chieu. 2012. *Optimising F-Measures: A Table of Tow Approaches*. Proceeding of The 29th International Conference on Machine Learning .Edinburgh. http://www.comp.nus.edu.sg/~leews/publications/fscore.pdf (accessed May. 22, 2013).

Evaluation Criteria of Information Retrieval Systems: What We Know and What We Do Not Know

Nadjla Hariri[1] | Fahime Babalhavaeji[2] | Mehrdad Farzandipour[3] | Somayyeh Nadi Ravandi[4]

1. Associate Professor, Department of Library and Information Science, Islamic Azad University, Tehran, Iran
nadjlahariri@gmail.co
2. Associate Professor, Department of Library and Information Science, Islamic Azad University. Tehran, Iran
f.babalhavaeji@gmail.com ac.ir
3. Associate Professor, Department of Health Information Technology, Kashan University of Medical Sciences, Kashan, Iran
farzandipour_m@kaums.ac.ir
4. [Corresponding Author] PhD Student, Department of Library and Information Science, Islamic Azad University, Tehran, Iran
nadi_so@kaums.ac.ir

Iranian Journal of Information Processing & Management

Iranian Research Institute for Science and Technology
ISSN 2251-8223
eISSN 2251-8231
Indexed in SCOPUS, ISC & LISA
Vol.30 | No.1 | pp: 199-221
Autumn 2014

Abstract: Evaluation of information retrieval systems is one of the greatest challenges for information science specialists, because determining the performance of a system depends on judgment of the relevance of documents provided by the system to the user's information needs, and it has its own complexities. New retrieval systems due to the dynamic nature of the Web are very different compared with traditional retrieval systems. Web information retrieval systems in terms of ranking results are divided into two groups: The ranked retrieval results and unranked retrieval sets. Each system has a different evaluation criteria and scales.

In this paper, criteria for evaluating information retrieval systems is reviewed for the ranked retrieval results (including precision and recall curves, interpolated precision, the 11-point interpolated precision average, Mean Average Precision, precision at K, R-precision, ROC curve cumulative gain and normalized discounted cumulative gain), and the unranked retrieval sets (including precision, recall, F-measure and accuracy) is introduced separately. Finally, alluding to the evaluation metrics, a scale of 4 degrees including best, useful, objective precision and then differential precision is introduced which can be used to

evaluate the non-binary precision of web information retrieval systems.

Keywords: Information Retrieval; Precision, Evaluation; Evaluation Criterion; Ranked Retrieval Results; Unranked Retrieval Sets