اصول و روش‌های نمایه‌سازی

جلسه پنجم: نمایه‌سازی در موتورهای کاوش و وب

هادی هراتی
دکتری علم اطلاعات و دانش‌شناسی

فاطمه ذاکری‌فرد
دانشجوی دکتری علم اطلاعات و دانش‌شناسی

مرکز اطلاع‌رسانی و کتابخانه مرکزی دانشگاه فردوسی مشهد
تابستان ۱۳۹۸

سرفصل جلسه پنجم:

- نمایه سازی در موتور های کاوش و وب
- ابزارهای کاوش در وب
- اجزای موتور جستجو
- رویکرد نمایه‌سازی در وب

نمایه‌سازی در موتورهای کاوش و وب:

- نمایه وب بر اساس روش‌های نمایه‌سازی ماشینی

- بررسی روش‌ها و اقدامات انجام‌گرفته، جهت آوردن صـــفحات وب به‌قصد نمایه‌سازی

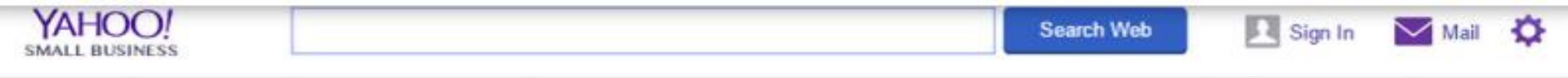
ابزارهای کاوش در وب:

- **Directories** (Human powered)
- **Search Engines** (Crawler-Based)
- **Meta Search Engines** (Mixed results)

Directories:

- نمایه اطلاعات صفحات یا سایت‌های وب منتخب توسط نیروی انسانی در پایگاه – عمومی یا خاص

- دسته‌بندی موضوعی اطلاعات به منظور دسترسی سریع استفاده کننده به مطالب مرتبط



Business to Business
Communications and Networking, Manufacturing, Computers...

U.S. States
California, Michigan, Virginia...

Shopping and Services
Apparel, Travel and Transportation, Communication and Information Management...

International
Fiji, Russia, Canada...

Employment and Work
Careers and Jobs,

Business and Economy
Directories, Organizations, Classifieds...

Entertainment
Television Shows, Consumer Electronics, Comedy...

Finance and Investment
Chats and Forums, Socially Responsible Investing, Reference and Guides...

Science
Engineering, Agriculture, Energy...

Arts
Design Arts, Visual Arts, Humanities...

Recreation
Outdoors, Sports, Games...

Society and Culture
Home and Garden, Weddings, Religion and Spirituality...

Social Science
Linguistics and Human Languages, Political Science, Psychology...

Government
U.S. Government, Law,

Education
Distance Learning, K-12, Career and Vocational...

News and Media
Business, Television, Newspapers...

Reference
Phone Numbers and Addresses, Calendars, Quotations...

Medicine

CATEGORIES (What's This?)

- Acupuncture@
- Anatomy@
- Andrology (5)
- Anesthesiology (96)
- Audiology (33)
- Aviation and Aerospace Medicine (11)
- Bioethics@
- Biostatistics (10)
- Bloodless Medicine (4)
- Booksellers@
- Cardiology (73)
- Chiropractic (69)
- Clinical Chemistry@
- Clinical Trials (27)
- Clinical Virology (1)
- Commercial Software@
- Comparative Medicine (10)
- Conferences (38)
- Cosmetic and Plastic Surgery (43)
- Mountain Medicine (8)
- Museums and Exhibits (14)
- Narrative Medicine (2)
- Nephrology (19)
- Neurology (107)
- Nobel Prize in Physiology or Medicine@
- Nuclear Medicine (14)
- Obstetrics and Gynecology (56)
- Occupational Medicine (11)
- Occupational Therapy (24)
- Oncology (67)
- Ophthalmology (82)
- Optometry (45)
- Organizations (489)
- Orthopedics (38)
- Osteopathy (39)
- Otolaryngology (31)
- Pain Management (25)
- Palliative Care (25)

Directories:

- **مزایا:**

1. کیفیت بهتر اطلاعات؛
2. دسترسی بهتر به اطلاعات مرتبط؛
3. صرف زمان کمتر برای دسترسی به اطلاعات مرتبط؛
4. سهولت مرور و بازیابی اطلاعات.

Directories:

- **محدودیت‌ها:**

1. پوشش کم اطلاعات موجود در وب؛
2. روزآمد نبودن اطلاعات؛
3. نیاز به آگاهی از ساختار سلسله مراتب موضوعی علوم.

Search Engines:

- ابزاری دارای برنامه‌های خودکار بدون وابستگی به نیروی انسانی
- فرایند شناسایی، انتخاب و نمایه‌سازی اطلاعات وب، توسط برنامه‌های رایانه‌ای

Search Engines:

- چگونه یک موتور جستجو در کمتر از یک ثانیه کل وب را جستجو و مطالب را برای ما بازیابی می‌کند؟

- جستجو و بازیابی مطالب از پایگاه اطلاعاتی خود که قبلاً از صفحات وب گردآوری و ذخیره نموده است.

Search Engines:

- اجزای موتور کاوش:

1) روبات (Robot)، خزنده (Crawler)، عنکبوت(Spider)

2) نمایه کننده (Indexer)

3) پایگاه اطلاعاتی (Database)

4) نرم‌افزار بازیابی اطلاعات (Search engine software)

۱. روبات‌ها:

- برنامه‌های خودکاری
- جستجو و شناسایی صفحات وب
- بر اساس ساختار فراپیوندی وب
- به طور پیوسته و در فواصل زمانی معین

شيوه كار روبات (كراولر):

1. آوردن صفحه URL هاي مشخص شده تحت عنوان seed page (Fetch)

2. صفحه آورده شده جهت استنتاج و استخراج متن و لينک ها تجزيه مي شوند (parsing)

3. متن استخراج شده به نمايه ساز متني وارد مي شود.

4. لينک هاي استنتاج شده به URL Frontier اضافه مي شود.

شیوه کار روبات (کراولر):

URL Frontier:

لینک‌هاي که هنوز خزنده صفحات مربوط به آنها را نیاورده است.

حرکت روبات (کراولر):

1. حرکت عمق- شروع

2. حرکت توزیع- شروع

3. حرکت بهترین- شروع. استفاده از سیاست رتبه‌بندي صفحات

حرکت روبات (کراولر):

سیاست رتبه بندي صفحات بر اساس:

الف) تعداد لینک‌هاي برقرارشده به آن صفحه

ب) افزایش اعتبار آن صفحه در صورتی که لینک‌های برقرار شده از صفحات معتبر باشد

ج) کاهش ارزش آن صفحه در صورت داشتن تعداد لینک‌های زیاد به سایر صفحات

د) تعداد بازدیدها از صفحه مورد نظر

۲. نمایه کننده:

- متون استخراج‌شده توسط خزنده‌ها در اختیار Indexer قرار می‌گیرند و آغاز فرایند نمایه‌سازی ماشینی

- تجزیه‌وتحلیل اطلاعات: اطلاعات از کدام صفحه ارسال‌شده است، چه حجمی دارد، کلمات موجود در آن کدام است، چند بار تکرار شده است، در کجای صفحه قرار دارند و ... .

- استفاده سیاهه بازدارنده

۳. پایگاه اطلاعاتی:

- ارسال اطلاعات تجزیه‌وتحلیل و نمایه‌سازی شده Indexer به پایگاه اطلاعاتی موتور جستجو

- انباری از اطلاعات سازمان‌دهی شده

- نمایش اطلاعات مورد جستجوی کاربر از داخل پایگاه اطلاعاتی توسط موتور جستجو

پایگاه اطلاعاتی:

- بزرگی و به روز بودن پایگاه اطلاعاتی یک موتور جستجو یک امتیاز

- یکی از تفاوت‌های اصلی موتورهای جستجو در حجم پایگاه اطلاعاتی آنها و همچنین روش ذخیره‌سازی داده‌ها است.

۴. نرم‌افزار بازیابی اطلاعات:

- واسط بین کاربر و پایگاه اطاعاتی
- از طریق واردکردن کلیدواژه‌ها در فیلدهای مختلف،
- جستجو در میلیون‌ها صفحه وب نمایه شده در پایگاه اطلاعاتی موتورهای کاوش
- رتبه‌بندی نتایج بازایابی براساس میزان تناسب و ارتباط آن با درخواست، واژه یا عبارت مورد نظر با استفاده از الگوریتم‌ها

نرم‌افزار بازیابی اطلاعات:

- اهمیت طراحی نرم‌افزار بازیابی اطلاعات از جنبه‌ها مختلف ازجمله رابط کاربری، روش‌های مختلف جستجو، رتبه‌بندی نتایج و .... چرا؟

- وجود پایگاه اطلاعاتی بسیار غنی در موتور جستجو ولی عدم توانایی نرم‌افزار بازیابی اطلاعات در بازیابی و ارائه نتایج

نرم‌افزار بازیابی اطلاعات:

معمولاً برای رتبه‌بندی دو ویژگی مهم در نظر گرفته می‌شود:

- **ویژگی‌های دورنی**: (محل درج کلید واژه، تعداد تکرار یا بسامد)

- **ویژگی‌های بیرونی**: (تعداد لینک به سایت مربوطه، تعداد بازدید از سایت مربوطه)

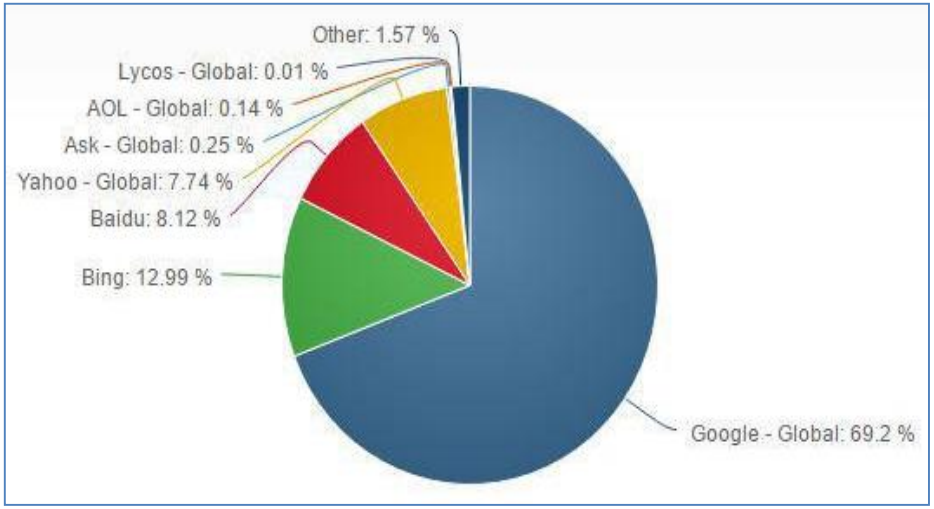
نرم‌افزار بازیابی اطلاعات:

نکته های آماری در خصوص اهمیت رتبه بندی:

- از نظر ۸۲ درصد کاربران اینترنت، موتورهای جستجو ابزار اصلی یافتن وب سایت ها
- از نظر بسیاری از کاربران در اغلب موارد ۱۰ رتبه اول نتایج جستجو، برآورده کننده خواسته آن‌ها
- ۷۰ درصد کاربران به هنگام جستجو یکی از سه سایت ابتدای نتایج جستجو را کلیک می کنند.
- فقط ۷ درصد سایت های قرار گرفته در رتبه های بیستم به بعد را کلیک می کنند.
- ۸۵ درصد از آن‌ها اگر جوابی برای جستجوی خود در بیست نتیجه اول نتایج جستجو نیابند، یا استراتژی جستجوی خود را تغییر می‌دهند یا موتور جستجو مورداستفاده خود را عوض می‌کنند.
- ۳۳ درصد کاربران وب فکر می‌کنند سایتی که در ابتدای نتایج جستجو قرار دارد، در موضوع جستجو شده سرآمدتر از رقبای خود است و در آن موضوع پیشرو است.

نرم‌افزار بازیابی اطلاعات:

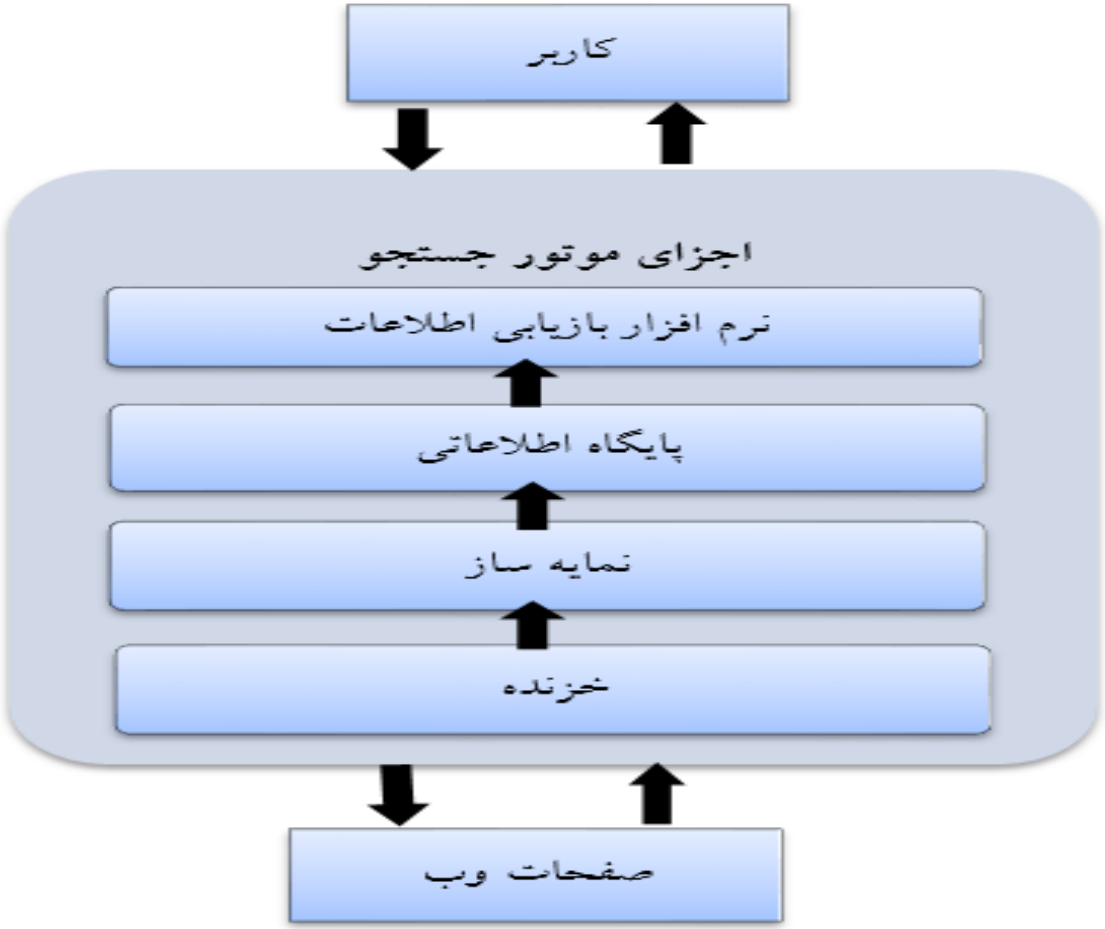
- این آمارها نشان‌دهنده اهمیت نرم‌افزار بازیابی اطلاعات و نظام رتبه‌بندی آن است. اگر موتور جستجو نتواند مطالب مرتبط با نیاز اطلاعاتی کاربر را بازیابی کند، بخش قابل‌توجهی از مخاطبان خود را از دست خواهد داد.



- موتور جستجوی گوگل

با الگوریتم‌های پیشرفته‌ای که دارد بخش اعظمی از مخاطبین را به خود جذب کرده است.

اجزای موتورهای کاوش:



Meta Search Engines:

- پوشش چندین موتور کاوش
- فاقد پایگاه مستقل

www.savvysearch.com/search/information+retrieval?q=information+retrieval&p=4

information retrieval

SavvySearch
CBS network

Select Domain

Search

9,710,000 results for: **information retrieval** in (0.19 seconds)

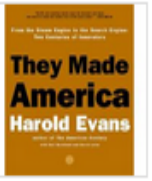
Filter Data

Sorting

New AI could help write your next textbook, using Penn State
www.techrepublic.com

Oct 26, 2015 He first came up with the idea when he was teaching Information Retrieval and Search Engines and wanted to build course material for the

2 0



Top metacrawler search engine deals at mySimon | Compare Search
www.mysimon.com

Search Engines: Information Retrieval in Practice - Bruce Croft - 9780136072249 - 0136072240. $136.95 Textbooks.com · They Made America: From the Steam

0 0

Top Search Engines List deals at mySimon | Compare Search

Top Search Engines List deals at mySimon | Compare Search
www.mysimon.com

Search Engines: Information Retrieval in Practice - New. Search Engines: Information Retrieval in Practice - Bruce Croft - 9780136072249 - 0136072240.

0 0

Related Searches

- **Information Retrieval**
- Information Retrieval Methods
- Information Retrieval System
- Information Retrieval Tutorial
- Information Retrieval Papers
- Information Storage And Retrieval
- Information Retrieval Theory
- Define Information Retrieval System

Featured Authors

Hope Reese

Latest News


metacrawler
Web
News
Images
Videos

meta crawler

Web News Images Videos

Library & **Information** Sciences - University of Illinois Masters

Ad online.illinois.edu/MSIM ▾

Join the #1 Program for Analyzing & Managing **Information**. Apply by Dec 1! Customizable Curriculum. Checkout Prices. Options: Online, On Campus.

▶ Visit Website

Free Full PDF Articles

Ad www.freefullpdf.com/ ▾

Over 80 million free publications Biology - Physics - Humanities. Qualified Experts. Sign Up For Mailing List. Founded In 2009. Highlights: Increase Visibility And Ease Of Use, Provide Customized Watching Services.

▶ Visit Website

prof. ingénieure et doctorante - niveau secondaire supérieur

Ad www.class-success.net/ ▾ +33 7 56 95 47 93

Ingénieure en sciences nucléaires, doctorante plasma, exp.18 ans vous prépare à la rentrée. Remise à niveau cet été en maths et sciences pour le secondaire et supérieur. remise à niveau maths. elearning-coaching. maths-sciences superieur. prof. maths sciences. Cours: cours particuliers, coaching maths.sciences, secondaire superieur.

▶ Visit Website

The 7 types of matter (PDF) - Includes the 7 substates - Solid, liquid, gas & 4 ether's

Ad www.alliancesforhumanity.com/matter ▾

This is the PDF version of the webpage AlliancesForHumanity.com/matter/ Download AlliancesForHumanity.com/matter/Matter.pdf if not automatically redirected.

▶ Visit Website

Searches related to **information sceince**

library and information science **careers**

information sciences **journal**

information science **degree jobs**

information science **and technology**

information sciences **theory**

what is information science **degree**

information science **umd**

information science **jobs**

Meta Search Engines:

مزایا:

- افزایش جامعیت در بازیابی اطلاعات
- صرفه جویی در زمان جستجو
- حذف نتایج تکراری

Meta Search Engines:

محدودیت ها:

- بازیابی بیش از حد اطلاعات
- کاهش مانعیت یا دقت در بازیابی اطلاعات
- جستجوی سطحی در پایگاه های اطلاعاتی

رویکردهای نمایه‌سازی خودکار در وب:

- محتوا محوری
- معنا محوری

محتوا محوری:

- استفاده از روش نمایه‌سازی برمبنای کلیدواژه‌های متن
- فرایند نمایه‌سازی شامل سه مرحله:

❖شکستن کلمات

❖ تعدیل و حذف کلمات غیرموضوعی

❖استفاده از الگوریتم ریشه‌ساز جهت تولید ریشه‌های مفاهیم

معنا محوری:

- توجه به مفاهیم، الگوها و کلیدهایی که به فهم مفاهیم می‌انجامد
- جست‌وجوی صرفا بر مبنای کلیدواژه‌ها نیست
- استفاده از فهرست مترادف‌ها
- بهره‌گیری از جست‌وجوی فازی

معنا محوری:

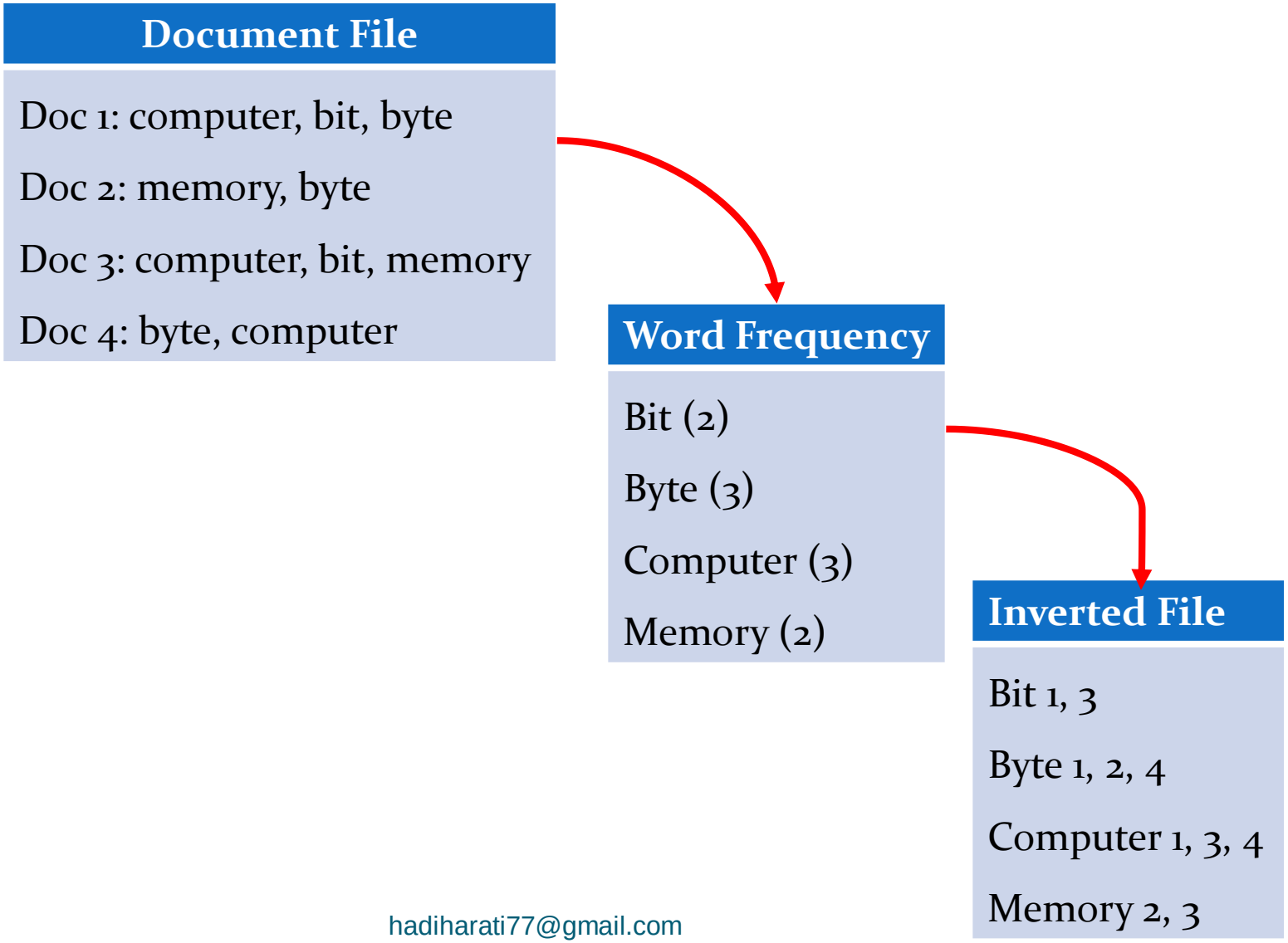
- بهره‌گیری از نمایه‌سازی معنـایی پنهـان جهـت بهبـود مانعیـت، جامعیت و رتبه‌بندی نتایج کاوش
- مهم‌ترین فایده‌ی وب معنایی تحول در نتایج بازیابی اطلاعات

چگونگی بازیابی اطلاعات:

- **Stop list**
- **Terms weighting**
- **CF : Collection Frequency**
- **DF: Document Frequency**

Words	CF	DF
Try	16422	8760
Insurance	19440	3997

Inverted Index:



تکنیک N-gram

- تکنیکی برای کاوش فازی
- در گوگل عبارت Did you mean:

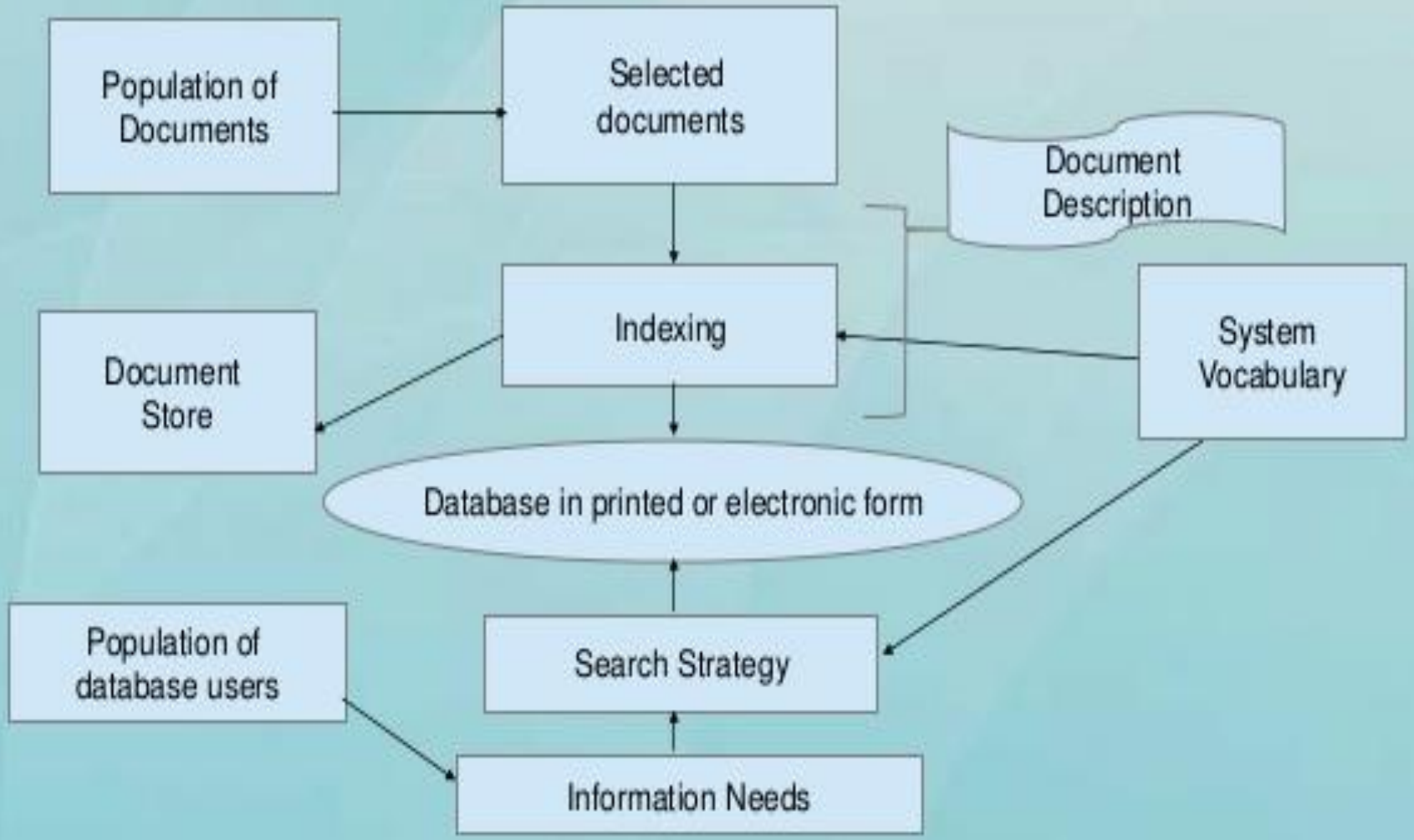
- Bi-gram (n=2)
- Tri-gram (n=3)
- Example:

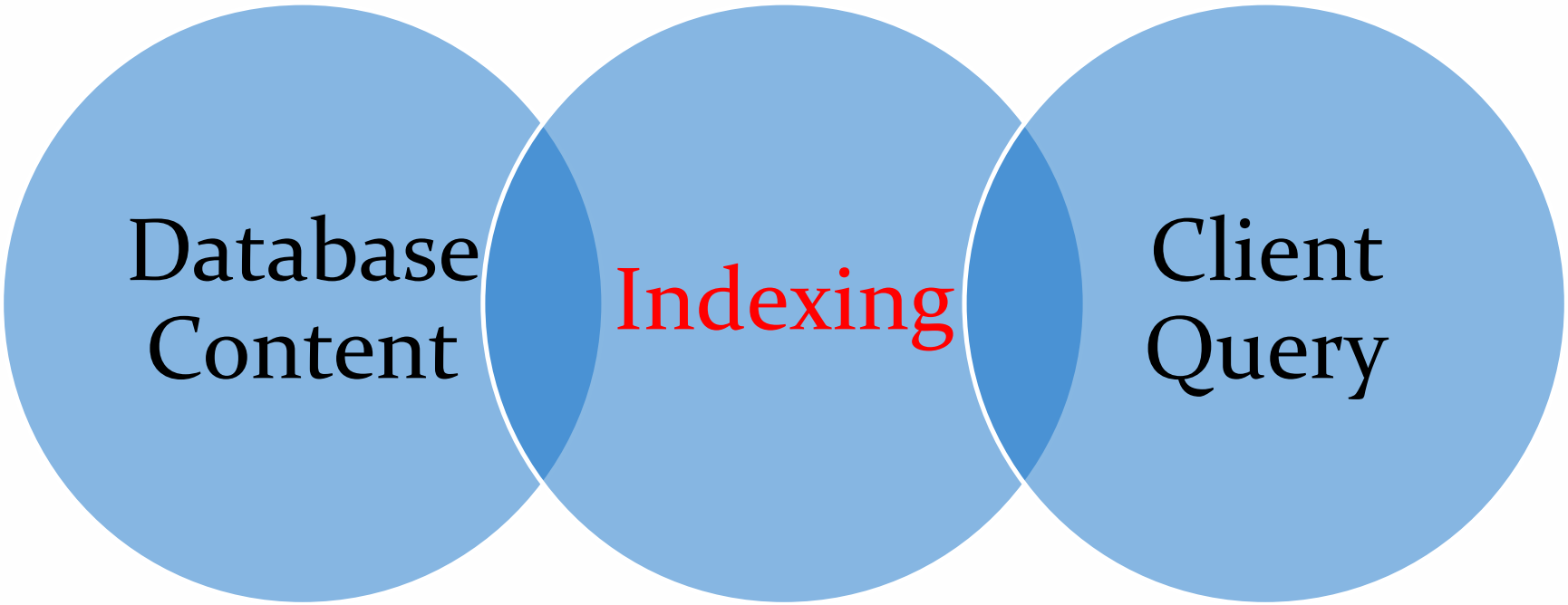
Compuiter & Computer

- Bi-gram: co om mp pu ui it te er
- Bi-gram: co om mp pu ut te er

- Tri-gram: com omp mpu pui uit ite ter
- Tri-gram: com omp mpu put ute ter

Role of Indexing in Information Retrieval




Database Content
Indexing
Client Query

آیا انتظاری که از این کارگاه داشتید برآورده شد؟

Thank you for your attention