المحتويات

[الطلب الأول 2](#_Toc513983237)

[الطلب الثاني 4](#_Toc513983238)

[الطلب الثالث 6](#_Toc513983239)

[1. Boolean model 6](#_Toc513983240)

[2. VECTOR SPACE MODEL 8](#_Toc513983241)

[3. Extended Boolean model 11](#_Toc513983242)

[الطلب الرابع 12](#_Toc513983243)

[الطلب الخامس 13](#_Toc513983244)

[الطلب السادس 14](#_Toc513983245)

[ الوثائق المهيكلة 14](#_Toc513983246)

[ الوثائق الغير مهيكلة 14](#_Toc513983247)

[ الوثائق شبه المهيكلة 15](#_Toc513983248)

# الطلب الأول

محركات البحث الخاصة بمجال معين فقط تقوم بفهرسة المستندات ذات الصلة بمجال معين ، مثل المعلومات الصحية أو معلومات الملكية الفكرية. بما أن المعرفة السابقة متوفرة حول مجال الاهتمام ، يمكن تكييف محركات البحث هذه للاستفادة من هذه المعرفة لتحسين نتائج البحث.

إن مهمة استرجاع المعلومات المخصصة هو ، العثور على الوثائق داخل مجموعة مثل الكتاب المقدس ، والتي تكون ذات صلة للمستخدم لا يزال يمثل تحدياً صعباً. في بعض الأحيان قد لا تحتوي المستندات ذات الصلة على الكلمة الأساسية المحددة. لا يعني عدم وجود المصطلح في وثيقة ما بالضرورة أن المستند ليس ذا صلة. لأن أكثر من مصطلح واحد يمكن أن يكون متشابهاً معًا على الرغم من اختلافه اللغوي.

في سياقات محددة ، قد تحتاج قائمة stop word للتوسع. وبالتالي فإن "المخدرات Drug" ستكون كلمة توقف (stop word) في مجموعة متعلقة بعلم الأدوية.

قد يكون قاموس المرادفات للأغراض العامة (على سبيل المثال ، تطبيق اللغة الإنجليزية) أو المجال المحدد. يعد نظام اللغة الطبية الموحد في المكتبة الوطنية للطب (UMLS) (وهو عبارة عن معجم أو تكتل لمحتويات العديد من المعاجم الطبية الحيوية) أكبر قاموس عالمي خاص بمجال محدد. أكبر وأفضل تركيبة من مفردات المكون الفردي داخل UMLS هي رؤوس الموضوعات الطبية (MeSH) المطورة في NLM لفهرس Medicus.

MeSH: Medical Subject Headings

* أنشأتها المكتبة الوطنية للطب لفهرسة المجلات الطبية الحيوية والكتب
* حوالي 25،000 عنوان الموضوعات مرتبة في التسلسل الهرمي (بشكل هرمي)
* يمكن أن يظهر العنوان في مواقع متعددة في التسلسل الهرمي.

هناك العديد من الفوائد من استرجاع المعلومات المخصصة ، وتنعكس هذه الفوائد في حقيقة أن أساليب الاسترجاع الأكثر شعبية كلها مخصصة. نظام استرجاع مخصص سريع البناء. يمكنك التوصل إلى طريقة جديدة تعتقد أنها يمكن أن تحسن استرجاعها ، وقمت بإضافتها. العديد من أساليب الاسترداد المخصصة أيضاً سريعة جداً ، تحتاج فقط إلى النظر في حدوث شروط الاستعلام في المستندات التي تظهر في وقت الاستعلام. وعلى الرغم من أنك قد تشكك في أداء محرك البحث المفضل على الإنترنت في بعض الأحيان ، فإن جودة نتائج البحث مرتفعة جداً في الواقع أيضاً.

# الطلب الثاني

تم تخصيص وقت وجهد كبير لتقليل الوقت بين استلام الاستعلام واستجابة محرك البحث، ولسبب وجيه. تشير الأبحاث إلى أنه حتى في حالات استرجاع أعلى قليلاً من قبل محركات بحث الويب يمكن أن تؤدي إلى انخفاض كبير في تصورات المستخدمين لجودة النتائج والتفاعل مع نتائج البحث. على الرغم من أن المستخدمين قد توقعوا استجابات سريعة من محركات البحث، فإن التطورات الحديثة في فهمنا للكيفية التي يعثر بها الأشخاص على المعلومات تشير إلى وجود سيناريوهات حيث يمكن لمحرك البحث أن يستغرق وقتاً أطول بكثير من جزء من الثانية لعرض محتوى ذي صلة.

على الرغم من أننا نعيش في عالم حيث كل شيء \_من الاتصالات لمعالجة المعلومات إلى وسائل النقل\_ هو الحصول على السرعة بشكل متزايد، في السنوات الأخيرة كان هناك عدد من "الحركات البطيئة" التي تدعو إلى إبطاء السرعة في مقابل اتخاذ إجراءات لتحسين الجودة. وربما تكون حركة الطعام البطيئة هي الحركة الأكثر شهرة، حيث يشجع البعض الأساليب التقليدية والمتنوعة لإعداد وجبات الطعام. ومع ذلك ، فقد توسعت الفكرة لتشمل السفر البطيء، والتكنولوجيا البطيئة وحتى العلوم بطيئة. بناءً على هذه الحركات البطيئة المختلفة، ناقشنا في هذه الورقة البحث البطيء ، حيث يتم استخدام وقت إضافي لتزويد الباحثين بتجربة بحث أعلى جودة مما يمكن تحقيقه بالنظر إلى القيود الزمنية الصارمة التقليدية. في هذه الورقة نركز على كيفية استخدام محركات البحث لوقت إضافي من أجل الحصول على نتائج أفضل، ولكن لاحظ أنه يمكن أيضاً تصميم تجربة البحث لتشجيع الأشخاص على استخدام وقت إضافي للتعبير عن موضوع بحثهم والتعرف عليه.

لتحقيق أوقات استجابة شبه فورية ، تقدم محركات البحث عددًا من التنازلات لتقليل التكاليف الحسابية، وبالتالي التضحية بالمكاسب المحتملة ذات الصلة لصالح الاستجابة السريعة لاستعلام المستخدم. على سبيل المثال ، فإنها تحد من تعقيد الميزات أو النماذج المستخدمة لتحديد الوثائق ذات الصلة من خلال تقديم افتراضات مبسطة للغاية عن اللغة ، مثل معالجة النص على أنه "حقيبة كلماتbag of words " غير مرتبة. يتجاهل المطابقة السريعة والموجهة إلى الكلمات الدلالات الغنية النص ولكن طريقة فعالة لالتقاط بعض التشابه الفعال بين الاستعلامات والمستندات. كما أنها تحد من مجموعة المستندات التي تم البحث عنها لأي استعلام محدد باستخدام أساليب مثل البحث عن التخزين المؤقت للبحث والفهرسة ، على الرغم من أن هذا قد يعني فقدان المحتوى ذي الصلة ، وتكبد تكاليف إضافية للبنية الأساسية.

تم استهداف أوقات الاستجابة السريعة بواسطة محركات البحث لسبب وجيه (جيد). تقترح الأبحاث أنه عندما تكون محركات بحث الويب أبطأ قليلاً في إرجاع نتائج البحث عنها بشكل طبيعي، فإن التأخير يؤدي إلى انخفاض كبير في تفاعل المستخدمين. على الرغم من اعتياد الباحثين على الردود شبه الفورية على استفساراتهم، إلا أن التطورات الحديثة في فهمنا للكيفية التي يعثر بها الأشخاص على المعلومات تشير إلى وجود بعض السيناريوهات حيث يمكن لمحرك البحث أن يستغرق وقتًا أطول بكثير من جزء من الثانية لتحديد المعلومات ذات الصلة وعرضها للمستخدمين.

قد يكون البحث البطيء مفيداً أيضاً في دعم المهام التي تمتد عبر جلسات متعددة أو المتابعة عبر أجهزة متعددة. في الوقت الذي تعمل فيه محركات البحث على تحسين قدرتها على التنبؤ بما إذا كان الشخص سيعود إلى مهمة ما ومتى يتم ذلك، يمكن استخدام الوقت بين المهام لمراقبة المعلومات التي تم عرضها من قبل للتغيير أو تحديد معلومات جديدة وتلخيصها، وبالتالي تحسين تجربة المستخدم عند البحث.

نوضح أن هناك حالات يبدو أن السرعة فيها أقل أهمية من غيرها، وتناقش السيناريوهات التي قد تتمكن فيها محركات البحث من الاستفادة من وقت إضافي لتوفير تجربة بحث أفضل بدلاً من تحقيق نتائج أسرع.

# الطلب الثالث

## Boolean model

Boolean queries: exact match

**النموذج البولياني** يعتبر من النماذج التقليدية، فهو أول نموذج مستخدم في نظم استرجاع المعلومات ويعتبر من أكثر النماذج التي تم الاعتماد عليها. وهو يستخدم من قبل جميع أنظمة استرجاع المعلومات التجارية اليوم. النتائج المسترجعة في المطابقة التامة هي مجموعة من الوثائق (بدون ترتيب(.

lots of results ➔ the query is too broad

no results➔ the query is too narrow

* إيجابيات النموذج البولياني:

1. سهل التطبيق
2. يصاغ الاستعلام بشكل واضح ومعبر عن طلب المستخدم بدقة.

* سلبيات النموذج البولياني:

1. يعتمد على المطابقة التامة بين كلمات الاستعلام وكلمات الوثائق وهذا يؤدي إلى استرجاع عدد قليل جداً من الوثائق أو على العكس تماماً.
2. من الصعب القيام بعملية ترتيب لنتائج البحث، مع أن بعض الوثائق المسترجعة قد تكون أهم من غيرها ولكن لا تأخذ ترتيب أعلى.
3. قد يجد المستخدم صعوبة في صياغة الاستعلام بالشكل المطلوب.
4. جميع كلمات الفهرسة لها أوزان متساوية.
5. يعتبره البعض أنه أقرب إلى [استرجاع البيانات](https://www.marefa.org/index.php?title=%D8%A7%D8%B3%D8%AA%D8%B1%D8%AC%D8%A7%D8%B9_%D8%A7%D9%84%D8%A8%D9%8A%D8%A7%D9%86%D8%A7%D8%AA&action=edit&redlink=1) (Data Retrieval) منه إلى [استرجاع المعلومات](https://www.marefa.org/%D8%A7%D8%B3%D8%AA%D8%B1%D8%AC%D8%A7%D8%B9_%D8%A7%D9%84%D9%85%D8%B9%D9%84%D9%88%D9%85%D8%A7%D8%AA) (Information Retrieval).

* Boolean Queries Example

User information need: Interested to know about **Everest** and **Nepal**

User Boolean query: **Everest** AND **Nepal**

Example of Input collection

Doc1= English tutorial and fast track

Doc2 = learning latent semantic indexing

Doc3 = Book on semantic indexing

Doc4 = Advance in structure and semantic indexing

Doc5 = Analysis of latent structures

**Query problem:** advance and structure AND NOT analysis

##### **Boolean Model Index Construction**

First we build the term-document incidence matrix which represents a list of all the distinct terms and their presence on each document (incidence vector). If the document contains the term than incidence vector is 1 otherwise 0.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Terms/doc** | **Doc1** | **Doc2** | **Doc3** | **Doc4** | **Doc5** |
| English | 1 | 0 | 0 | 0 | 0 |
| Tutorial | 1 | 0 | 0 | 0 | 0 |
| Fast | 1 | 0 | 0 | 0 | 0 |
| Track | 1 | 0 | 0 | 0 | 0 |
| Books | 0 | 1 | 0 | 0 | 0 |
| Semantic | 0 | 1 | 1 | 1 | 0 |
| Analysis | 0 | 1 | 0 | 0 | 1 |
| Learning | 0 | 0 | 1 | 0 | 0 |
| Latent | 0 | 0 | 1 | 0 | 1 |
| Indexing | 0 | 0 | 1 | 1 | 0 |
| Advance | 0 | 0 | 0 | 1 | 0 |
| Structures | 0 | 0 | 0 | 1 | 1 |

So now we have 0/1 vector for each term. To answer the query we take the vectors for **advance**, **structure** and **analysis,** complement the last, and then do a bitwise AND.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Doc1** | **Doc2** | **Doc3** | **Doc4** | **Doc5** |  |
| 0 | 0 | 0 | 1 | 0 |  |
| 0 | 0 | 0 | 1 | 1 | (AND) |
| 0 | 0 | 0 | 1 | 0 |  |
| 1 | 0 | 1 | 1 | 0 | (NOT analysis) |
| 0 | 0 | 0 | 1 | 0 |  |
|  |  |  | Doc4 |  |  |

Hence Doc4 is retrieved here.

## VECTOR SPACE MODEL

Documents and queries are both vectors.

يتم استخدامه في تصفية واسترجاع المعلومات، كان أول استخدام لها في نظام استرجاع المعلومات SMART.

يقترح إطاراً يسمح بالمطابقة الجزئية، فكرة النموذج هي إرفاق أوزان الاستعلام عن المصطلحات وتمثيل المستندات والاستعلامات كمتجهات وحساب درجة التشابه بين تلك المتجهات. إن إحداثيات المتجهات هي أوزان مصطلحات الاستعلام في الوثيقة (لمتجه المستند) والاستعلام. إذا احتوى الاستعلام على مصطلحات v ، فإن المستندات والاستعلام يتم تمثيلهما في فضاء متجه الأبعاد. من المفترض عموماً أن تكون المصطلحات مستقلة.

هذا ما يفعله نظام NCBI’s Entrez الذي يصل إلى عدة قواعد بيانات، بما في ذلك (PubMed) تتم مقارنة كل وثيقة مع كل مستند آخر تم تحميله مسبقاً في قاعدة البيانات، باستخدام خوارزمية صممها John Wilbur من NCBI ، ويتم تخزين أعلى المطابقات لاسترجاعها في المستقبل.

* الإيجابيات

1. Simple model based on linear algebra;
2. Term weights not binary;
3. Allows computing a continuous degree of similarity between queries and documents;
4. Allows ranking documents according to their possible relevance;
5. Allows partial matching.

* Example

لنفترض لدينا مجموعة الوثائق التالية :

d1: “new york times”

d2: “new york post”

d3: “los angeles times”

نلاحظ ظهور بعض المصطلحات في مستندين، بينما البعض الآخر في مستند واحد فقط.

أولاً نقوم بحساب قيمة تكرار كل مصطلح (term frequency) فنجد:

Tf (new) = 2;

Tf (york) = 2;

Tf (times) = 2;

Tf (post) = 1;

Tf (los)= 1;

Tf (angeles) = 1;

نقوم بحساب قيمة idf حيث العدد الإجمالي للوثائق هو N = 3 :

Idf(new) = log2(N/tf) = log2(3/2)=0.584

Idf(york) = log2(3/2)=0.584

idf(times) = log2(3/2)=0.584

idf(post) = log2(3/1)=1.584

Idf(los) = log2(3/1)=1.584

Idf(angles) = log2(3/1) = 1.584

بفرض تم الترتيب أبجدياً:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| york | times | post | new | los | angles |  |
| 1 | 1 | 0 | 1 | 0 | 0 | D1 |
| 1 | 0 | 1 | 1 | 0 | 0 | D2 |
| 0 | 1 | 0 | 0 | 1 | 1 | D3 |

الآن نقوم بحساب tf-idf من خلال جداء كل من idf و tf كالتالي:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| york | times | post | new | los | angles |  |
| 0.584 | 0.584 | 0 | 0.584 | 0 | 0 | D1 |
| 0.584 | 0 | 1.584 | 0.584 | 0 | 0 | D2 |
| 0 | 0.584 | 0 | 0 | 1.584 | 1.584 | D3 |

على اعتبار أن الاستعلام هو "new new times"، نقوم بحساب tf-idf له بنفس الطريقة السابقة فنحصل على :

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 0 | (1/2)\*0.584=0.292 | 0 | (2/2)\*0.584=0.584 | 0 | 0 | q |

ثم نقوم بحساب طويلة كل شعاع (الاستعلام والوثيقة) :

Length of d1 = sqrt(0.584^2+0.584^2+0.584^2)=1.011

Length of d2 = sqrt(0.584^2+1.584^2+0.584^2)=1.786

Length of d3 = sqrt(1.584^2+1.584^2+0.584^2)=2.316

Length of q = sqrt(0.584^2+0.292^2)=0.652

ومنه قيم التشابه هي:

cosSim(d1,q) = (0\*0+0\*0+0.584\*0.584+0\*0+0.584\*0.292+0.584\*0) / (1.011\*0.652) = 0.776

cosSim(d2,q) = (0\*0+0\*0+0.584\*0.584+1.584\*0+0\*0.292+0.584\*0) / (1.786\*0.652) = 0.292

cosSim(d3,q) = (1.584\*0+1.584\*0+0\*0.584+0\*0+0.584\*0.292+0\*0) / (2.316\*0.652) = 0.112

وفقاً لقيم التشابه، سيكون الترتيب النهائي الذي يتم عرض المستندات به كنتيجة للاستعلام هو: d1 ، d2 ، d3

## Extended Boolean model

الهدف من نموذج Boolean الموسع هو التغلب على عيوب النموذج Boolean الذي تم استخدامه في استرجاع المعلومات.النموذج البولياني لا يأخذ بعين الاعتبار الأوزان للمصطلحات في الاستعلامات ، وغالباً ما تكون مجموعة نتائج الاستعلام المنطقي صغيرة جداً أو كبيرة جداً.

تتمثل فكرة النموذج الموسع في الاستفادة من الأوزان الجزئية والموازنة كما في نموذج فضاء المتجه. فهو يجمع بين خصائص نموذج Vector Space وخصائص الجبر البولياني، ويصنف التشابه بين الاستعلامات والمستندات. بهذه الطريقة قد تكون الوثيقة ذات صلة إلى حد ما إذا كانت تتطابق مع بعض المصطلحات التي تم الاستعلام عنها وستتم إعادتها كنتيجة، بينما في نموذج Boolean القياسي لم تكن كذلك.

# الطلب الرابع

البعد الدلالي مهم جداً لنظام استرجاع المعلومات لأنه من الوارد جداً لو أردنا استرجاع بيانات هامة من صفحة أو وثيقة لسنا على علم بهيكليتها, ففي الحالة التقليدية لا يمكننا الوصول مباشرة إلى معلومات هامة ودقيقة ضمن صفحة أو وثيقة لا نعلم بهيكليتها مسبقاً.

ولكن بإضافة البعد الدلالي، أولاً يتم التنبؤ بمعنى ومرادف كل توصيف موجود بالصفحة, بغض النظر عن التعبير أو اللغة المكتوب بها، وربطه بمرادف آخر معروف بالنسبة للنظام, ثم يتم تحليل الصفحات ودراستها في مرحلة الاستعلام والوصول مباشرة إلى ما يطلبه المستخدم دون أن نكون على علم مسبق بهيكلية هذه الصفحات.

وبالنسبة للوثائق من خلاله نستطيع الاستفسار ليس فقط عن الكلمات المفهرسة في الكتاب وإنما أيضاً عما يتعلق بها كأن نستعلم باستخدام مرادف الكلمات، وبالتالي لا يتم حصر عملية الاستعلام في الكلمات الواردة في الوثائق فقط. أي الحصول على نتائج موائمة لطلب المستخدم بشكل أكثر بحيث تلبي حاجته من المعلومات.

نلاحظ في اللغات الطبيعية بنظام استرجاع المعلومات نواجه عدة مشاكل منها:

* Homonym : وهو وجود الكلمة ولها معنيين مختلفين.
* Polysemy : وهو وجود الكلمة ولها أكثر من معنى متشابه.
* Synonyms : وهو وجود أكثر من كلمة لها نفس المعنى.

ولذلك تعتبر الخطوات السابقة غير كافية ولذلك في البعد الدلالي نعتمد على مبدأ توسيع الـ query بالاعتماد على knowledge base ,مثال : Wordnet والتي تجمع المعرفة بلغات معينة ويتم ربط الاستعلام الذي أدخله المستخدم بالـknowledge base والتي هي الـ ontology.

ويتم إضافة البعد الدلالي وفقاً للخطوات التالية :

عند الوصول للـindex terms , نقوم بأخذها ونستبعد الأكثر تكراراً والأقل تكراراً وعديمة التواتر.

المرحلة الثانية نقوم بعمل reweighting (إعادة توزين ) لل index terms المتبقية، وبالتالي أصبح لدينا index terms جديدة نقوم بعمل Matching function عليها.

# الطلب الخامس

نعم يمكننا، أولاً يمكن إدخال البيانات المكانية إلى الأنتولوجي وهي تسمح بوصف المعرفة بطريقة معززة يمكن أن تستخدم لتنفيذ الاستدلال على البيانات المكانية.

في نموذج الاستمرار والذي يتم تمثيل الكائن بعدة جوانب على سبيل المثال الجانب الدلالي والجانب المكاني والجانب الزماني ويتم التعامل مع كل جانب بشكل مستقل, فكل تغيّر يولد كائن جديد بشكل أتوماتيكي فإذا حصل تغير على إحدى الجوانب وليكن الدلالي مثلاً يحافظ على جوانبه السابق ويغير الجانب الدلالي ضمن الكائن الجديد الذي تم توليده أتوماتيكياً.

# الطلب السادس

## الوثائق المهيكلة

وتتمثل بقواعد المعطيات حيث تسمح بالبحث عن البيانات المطابقة تماماً للاستعلام وكذلك البحث ضمن مجال محدد ويوجد العديد من نظم إدارة قواعد المعطيات والتي تسهل البحث ضمن البيانات المهيكلة بشكل كبير وتعتمد بشكل اساسي على وجود دلالات واضحة للبيانات والاستعلامات .

يتم استرداد المعلومات وفقاً لهيكل معين, لم يعد اعتبار المستندات ككيانات ذرية ولكن كمجاميع كائنات مترابطة يمكن استردادها بشكل منفصل, أي في حال تقديم استعلام استرجاع, يمكن للمرء استرداد مجموعة مكونات الوثيقة الأكثر ملائمة لهذا الاستعلام. أي في المحصلة يمكن إعادة جزء من ال document.

فالمبدأ الأساسي لنظم استرجاع المعلومات المهيلكة:

A system should always retrieve the most specific part of a document answering the query

إذن هو يعيد أصغر وحدة تحوي على المعلومات المطلوبة، وفي هذا النوع من نظم استرجاع المعلومات تكون ال queries أكثر تعقيداً وتطوراً. ويكون أكثر كفاءة أثناء التصفح أو الاستعلام عن معلومات.

## الوثائق الغير مهيكلة

وتتمثل بالنصوص الحرة وهي الأصعب بالتعامل حيث لا يمكن لأنظمة البحث في البيانات المهيلكة أن تتعامل معها, ويتم البحث فيها باستخدام استعلامات مبينية على الكلمات المفتاحية وكذلك استعلامات مفاهمية أعقد من الاستعلامات النصية, ومن مشاكل البيانات الحرة أنها غير متناسقة لأنها مكتوبة بلغة بشرية فيمكن أن تمثل نفس المعلومة بعدة طرق ويمكن لعدة صيغ أن تمثل نفس المعلومة.

يعيد كامل الـ document , ويبحث ضمن free text , ليس لديها لغة علنية واضحة وأكثر مثال يوضحها هو النص الحر على صفحات الويب والصوت والفيديو، في حين تكون الاستعلامات queries بسيطة للغاية حتى يتمكن الحاسوب من فهمها.

## الوثائق شبه المهيكلة

شكل من أشكال البيانات المنظمة (structure data) التي لا تتوافق مع البينية الرسمية لنماذج البيانات العلائقية أو الأشكال الاخرى لجداول البيانات ومع ذلك فهي تحوي على علامات لفصل العناصر الدلالية وتفرض التسلسلات الهرمية للسجلات والحقول داخل البيانات. مثال عليها صفحات الويب والتي يتم كتابة المحتوى فيها بطريقة شبه مهيكلة باستخدام أحد لغات توصيف النصوص(markup languages) وأشهرها في توصيف صفحات الويب هي الـ html التي يتم فيها هيكلة النصوص ضمن tags محددة لها صفات محددة وبالتالي إعطاء هذه النصوص بعداً توصيفيّاً ومعرفياً إلى حد ما, بعكس الوثائق التي تكون غير مهيكلة.

|  |  |  |
| --- | --- | --- |
| Structure data | Un-Structure data | Semi-Structure data |
| Refer to information in "tables" | refer to free text | Almost no data is "un-structure data" |
| Allows exact match for (text) queries | Allows keyword queries includes operators.  Classic model for searching text documents. | Facilitates search such as title contains data |