Search Engines--Information Retrieval in Practice

tanshuqiu

Email: tsq@cqut.edu.cn

1

Queries and Interfaces

First : Information Needs and Queries

Second : Query Transformation and Refinement

Third : Showing the Results

Fourth : Cross-Language Search

Information Needs and Queries

From the user’ s point of view the search engine is primarily an interface for specifying queries and examining results. People can interact with the system during query formulation and reformulation, and while they are browsing the results. These interactions are a crucial part of the process of information retrieval, and can determine whether the search engine is viewed as providing an effective service.

Information Needs and Queries

From the point of view of the search engine designer, there are two important consequences of these observations about information needs:

Queries can represent very different information needs and may require different search techniques and ranking algorithms to produce the best rankings.

Information Needs and Queries

A query can be a poor representation of the information need. This can happen because the user finds it difficult to express the information need. More often, however, it happens because the user is encouraged to enter short queries, both by the search engine interface and by the fact that long queries often fail.

Query Transformation and Refinement

- ✓Stopping and Stemming Revisited
- ✓Spell Checking and Suggestions
- ✓Query Expansion
- ✓Relevance Feedback

Stopping and Stemming Revisited

Query-based stemming is another technique for increasing the flexibility of the search engine. If the words in documents are stemmed during indexing, the words in the queries must also be stemmed.

Stopping and Stemming Revisited

The query “fish village” will, for example, produce very different results from the query “fishing village”, but many stemming algorithms would reduce “fishing” to “fish”. By not stemming during document indexing, we are able to make the decision at query time whether or not to stem “fishing”.

Stopping and Stemming Revisited

For query-based stemming to work, we must expand the query using the appropriate word variants, rather than reducing the query word to a word stem. If the query word “fishing” was replaced with the stem “fish”, the query would no longer match documents that contained “fishing”.

Stopping and Stemming Revisited

For example, here are three stem classes created with the Porter Stemmer on TREC news collections (the first entry in each list is the stem):

/bank banked banking bankings banks
/ocean oceaneering oceanic oceanics oceanization oceans
/polic polical polically police policeable policed
-policement policer policers polices policial
-policically policier policiers policies policing
-policization policize policly policy policying policys

Stopping and Stemming Revisited

/bank banked banking bankings banks
/ocean oceaneering oceanic oceanics oceanization oceans
/polic polical polically police policeable policed
-policement policer policers polices policial
-policically policier policiers policies policing
-policization policize policly policy policying policys

These classes are not only long (the "polic" class has 22 entries), but they also contain a number of errors. The words relating to "police" and "policy" should not be in the same class, and this will cause a loss in ranking accuracy.

Stopping and Stemming Revisited

/bank banked banking bankings banks
/ocean oceaneering oceanic oceanics oceanization oceans
/polic polical polically police policeable policed
-policement policer policers polices policial
-policically policier policiers policies policing
-policization policize policly policy policying policys

Adding 22 words to a simple query will certainly negatively impact response time and, if not done properly using a synonym operator, could cause the search to fail.

Stopping and Stemming Revisited

/bank banked banking bankings banks
/ocean oceaneering oceanic oceanics oceanization oceans
/polic polical polically police policeable policed
-policement policer policers polices policial
-policically policier policiers policies policing
-policization policize policly policy policying policys

Both of these issues can be addressed using an analysis of word co-occurrence in the collection of text. The term association measure used in TREC experiments was based on Dice’ s coefficient.

Stopping and Stemming Revisited

Applying this technique to the three example stem classes, and using TREC data to do the co-occurrence analysis, results in the following connected components:

/policies policy
/police policed policing
/bank banking banks

Spell Checking and Suggestions

Spell checking is an extremely important part of query processing. Approximately 10–15% of queries submitted to web search engines contain spelling errors, and people have come to rely on the “Did you mean: ...” feature to correct these errors.

Spell Checking and Suggestions

The basic approach used in many spelling checkers is to suggest corrections for words that are not found in the spelling dictionary. Suggestions are found by comparing the word that was not found in the dictionary to words that are in the dictionary using a similarity measure.

Spell Checking and Suggestions

The most common measure for comparing words (or more generally, strings) is the edit distance, which is the number of operations required to transform one of the words into the other.

Spell Checking and Suggestions

The DamerauLevenshtein distance metric counts the minimum number of insertions, deletions, substitutions, or transpositions of single characters required to do the transformation. Studies have shown that 80% or more of spelling errors are caused by an instance of one of these types of single-character errors.

Spell Checking and Suggestions

extenssions → extensions (insertion error)
poiner → pointer (deletion error)
marshmellow → marshmallow (substitution error)
brimingham → birmingham (transposition error)

As an example, the following transformations (shown with the type of error involved) all have Damerau-Levenshtein distance 1 since only a single operation or edit is required to produce the correct word:

Spell Checking and Suggestions

doceration → deceration

deceration → decoration

The transformation doceration → decoration, on the other hand, has edit distance 2 since it requires two edit operations:

Spell Checking and Suggestions

A variety of techniques and data structures have been used to speed up the calculation of edit distances between the misspelled word and the words in the dictionary. These include restricting the comparison to words that start with the same letter(since spelling errors rarely change the first letter), words that are of the same or similar length (since spelling errors rarely change the length of the word), and words that sound the same.

Spell Checking and Suggestions

The Soundex code is a simple type of phonetic encoding that was originally used for the problem of matching names in medical records. The rules for this encoding are:

1. Keep the first letter (in uppercase).
2. Replace these letters with hyphens: a, e, i, o, u, y, h, w.
3. Replace the other letters by numbers as follows:
    1: b, f, p, v
    2: c, g, j, k, q, s, x, z
    3: d, t
    4: l
    5: m, n
    6: r
4. Delete adjacent repeats of a number.
5. Delete the hyphens.
6. Keep the first three numbers or pad out with zeros.

Some examples of this code are:

extenssions → E235; extensions → E235

marshmellow → M625; marshmallow → M625

brimingham → B655; birmingham → B655

poiner → P560; pointer → P536

Query Expansion

The key to effective expansion is to choose words that are appropriate for the context, or topic, of the query. For example, “aquarium” may be a good expansion term for “tank” in the query “tropical fish tanks”, but not appropriate for the query “armor for tanks”.

Query Expansion

Term association measures are an important part of many approaches to query expansion, and consequently a number of alternatives have been suggested.

Measure	*Formula*
Mutual information (MIM)	$\frac{n_{ab}}{n_a . n_b}$
Expected Mutual Information ($EMIM$)	$n_{ab} . \log(N . \frac{n_{ab}}{n_a . n_b})$
Chi-square (χ^2)	$\frac{(n_{ab} - \frac{1}{N} . n_a . n_b)^2}{n_a . n_b}$
Dice's coefficient ($Dice$)	$\frac{n_{ab}}{n_a + n_b}$

MIM	*EMIM*	χ^2	*Dice*
trmm	forest	trmm	forest
itto	tree	itto	exotic
ortuno	rain	ortuno	timber
kuroshio	island	kuroshio	rain
ivirgarzama	like	ivirgarzama	banana
biofunction	fish	biofunction	deforestation
kapiolani	most	kapiolani	plantation
bstilla	water	bstilla	coconut
almagreb	fruit	almagreb	jungle
jackfruit	area	jackfruit	tree
adeo	world	adeo	rainforest
xishuangbanna	america	xishuangbanna	palm
frangipani	some	frangipani	hardwood
yuca	live	yuca	greenhouse
anthurium	plant	anthurium	logging

Most strongly associated words for "tropical" in a collection of TREC news stories. Co-occurrence counts are measured at the document level.

MIM	*EMIM*	χ^2	*Dice*
zoologico	water	arlsq	species
zapanta	species	happyman	wildlife
wrint	wildlife	outerlimit	fishery
wpfmc	fishery	sportk	water
weighout	sea	lingcod	fisherman
waterdog	fisherman	longfin	boat
longfin	boat	bontadelli	sea
veracruzana	area	sportfisher	habitat
ungutt	habitat	billfish	vessel
ulocentra	vessel	needlefish	marine
needlefish	marine	damaliscu	endanger
tunaboat	land	bontebok	conservation
tsolwana	river	taucher	river
olivacea	food	orangemouth	catch
motoroller	endanger	sheepshead	island

Most strongly associated words for "fish" in a collection of TREC news stories. Co-occurrence counts are measured at the document level.

Relevance Feedback

In relevance feedback the user indicates which documents are interesting(i.e., relevant) and possibly which documents are completely off-topic (i.e., non-relevant). Based on this information, the system automatically reformulates the query by adding terms and reweighting the original terms, and a new ranking is generated using this modified query.

Relevance Feedback

The specific method for modifying the query depends on the underlying retrieval model. In general, words that occur more frequently in the relevant documents than in the non-relevant documents, or in the collection as a whole, are added to the query or increased in weight.

Relevance Feedback

The same general idea is used in the technique of pseudo-relevance feedback, where instead of asking the user to identify relevant documents, the system simply assumes that the top-ranked documents are relevant.

Relevance Feedback

The same general idea is used in the technique of pseudo-relevance feedback, where instead of asking the user to identify relevant documents, the system simply assumes that the top-ranked documents are relevant.

Relevance Feedback

As a simple example of how this process works, consider the ranking shown in this figure, which was generated using a popular search engine with the query “tropical fish”. To expand this query using pseudo-relevance feedback, we might assume that all these top 10 documents were relevant.

1. Badmans **Tropical Fish**

   A freshwater aquarium page covering all aspects of the **tropical fish** hobby. ... to Badman's **Tropical Fish**. ... world of aquariology with Badman's **Tropical Fish**. ...

2. **Tropical Fish**

   Notes on a few species and a gallery of photos of African cichlids.

3. The **Tropical** Tank Homepage **- Tropical Fish** and Aquariums

   Info on **tropical fish** and **tropical** aquariums, large **fish** species index with ... Here you will find lots of information on **Tropical Fish** and Aquariums. ...

4. **Tropical Fish** Centre

   Offers a range of aquarium products, advice on choosing species, feeding, and health care, and a discussion board.

5. **Tropical fish** - Wikipedia, the free encyclopedia

   Tropical fish are popular aquarium **fish** , due to their often bright coloration. ... Practical Fishkeeping • **Tropical Fish** Hobbyist • Koi. Aquarium related companies: ...

6. **Tropical Fish** Find

   Home page for **Tropical Fish** Internet Directory ... stores, forums, clubs, **fish** facts, **tropical fish** compatibility and aquarium ...

7. Breeding **tropical fish**

   ... intrested in keeping and/or breeding **Tropical**, Marine, Pond and Coldwater **fish**. ... Breeding **Tropical Fish** ... breeding **tropical**, marine, coldwater & pond **fish**. ...

8. FishLore

   Includes **tropical** freshwater aquarium how-to guides, FAQs, **fish** profiles, articles, and forums.

Relevance Feedback

By analyzing the full text of these documents, the most frequent terms, with their frequencies, can be identified as:

a (926), td (535), href (495), http (357), width (345), com (343), nbsp (316), www (260), tr (239), htm (233), class (225), jpg (221)

Relevance Feedback

Clearly, these words are not appropriate to use as expansion terms, because they consist of stop words and HTML expressions that will be common in the whole collection. In other words, they do not represent the topics covered in the top ranked documents.

Relevance Feedback

A simple way to refine this process is to count words in the snippets of the documents and ignore stop words. This analysis produces the following list of frequent words:

tropical (26), fish (28), aquarium (8), freshwater (5), breeding (4), information (3), species (3), tank (2), Badman's (2), page (2), hobby (2), forums (2)

Context and Personalization

The query context, will affect the relevance of retrieved documents and could potentially have a significant impact on the ranking algorithm. Most contextual information, however, has proved to be difficult to capture and represent in a way that provides consistent effectiveness improvements.

Context and Personalization

There are examples of applications where the use of contextual information is clearly effective. One of these is the use of query logs and click through data to improve web search. The context in this case is the history of previous searches and search sessions that are the same or very similar.

Context and Personalization

Another effective application of context is local search, which uses geographic information derived from the query, or from the location of the device that the query comes from, to modify the ranking of search results.

Context and Personalization

For example, the query “fishing supplies” will generate a long list of web pages for suppliers from all over the country (or the world). The query “fishing supplies Cape Cod”, however, should use the context provided by the location “Cape Cod” to rank suppliers in that region higher than any others.

Showing the Results

- ✓Result Pages and Snippets
- ✓Advertising and Search
- ✓Clustering the Results

Result Pages and Snippets

For most search engines the result pages consist of a ranked list of document summaries that are linked to the actual documents or web pages. A document summary for a web search typically contains the title and URL of the web page, links to live and cached versions of the page, and, most importantly, a short text summary, or snippet, that is used to convey the content of the page.

Advertising and Search

Advertising is a key component of web search engines since that is how companies generate revenue. In the case of advertising presented with search results (sponsored search), the goal is to find advertisements that are appropriate for the query context.

Advertising and Search

Search engine companies maintain a database of advertisements, Which is searched to find the most relevant advertisements for a given query or web page. An advertisement in this database usually consists of a short text description and a link to a web page describing the product or service in more detail. Searching the advertisement database can therefore be considered a special case of general text search.

Advertising and Search

By taking all of These factors into account, namely relevance, bids, and popularity, the search engine company can devise strategies to maximize their expected profit.

Advertising and Search

As an example, a pet supplies company that specializes in tropical fish may place the highest bid for the keywords “aquarium” and “tropical fish”. Given the query “tropical fish”, this keyword is certainly relevant. The content of the advertisement for that company should also contain words that match the query. Given that, this company’s advertisement will receive a high score for relevance and a high score based on the bid.

Advertising and Search

When queries are compared to advertisements. Advertisements contain a small number of words or keywords relative to a typical page, and the database of advertisements will be several orders of magnitude smaller than the Web.

Advertising and Search

For example, if a pet supply company has placed a high bid for “aquarium”, they would expect to receive some traffic from queries about “fish tanks”. This, of course, is the classic vocabulary mismatch problem, and many techniques have been proposed to address this, such as stemming and query expansion. Since advertisements are short, techniques for expanding the documents as well as the queries have been considered.

Advertising and Search

Two techniques that have performed well in experiments are query reformulation based on user sessions in query logs and expansion of queries and documents using external sources, such as the Web.

Advertising and Search

Studies have shown that about 50% of the queries in a single session are reformulations, where the user modifies the original query through word replacements, insertions, and deletions. Given a large number of candidate associations between queries and phrases in those queries, statistical tests, can be used to determine which associations are significant.

Advertising and Search

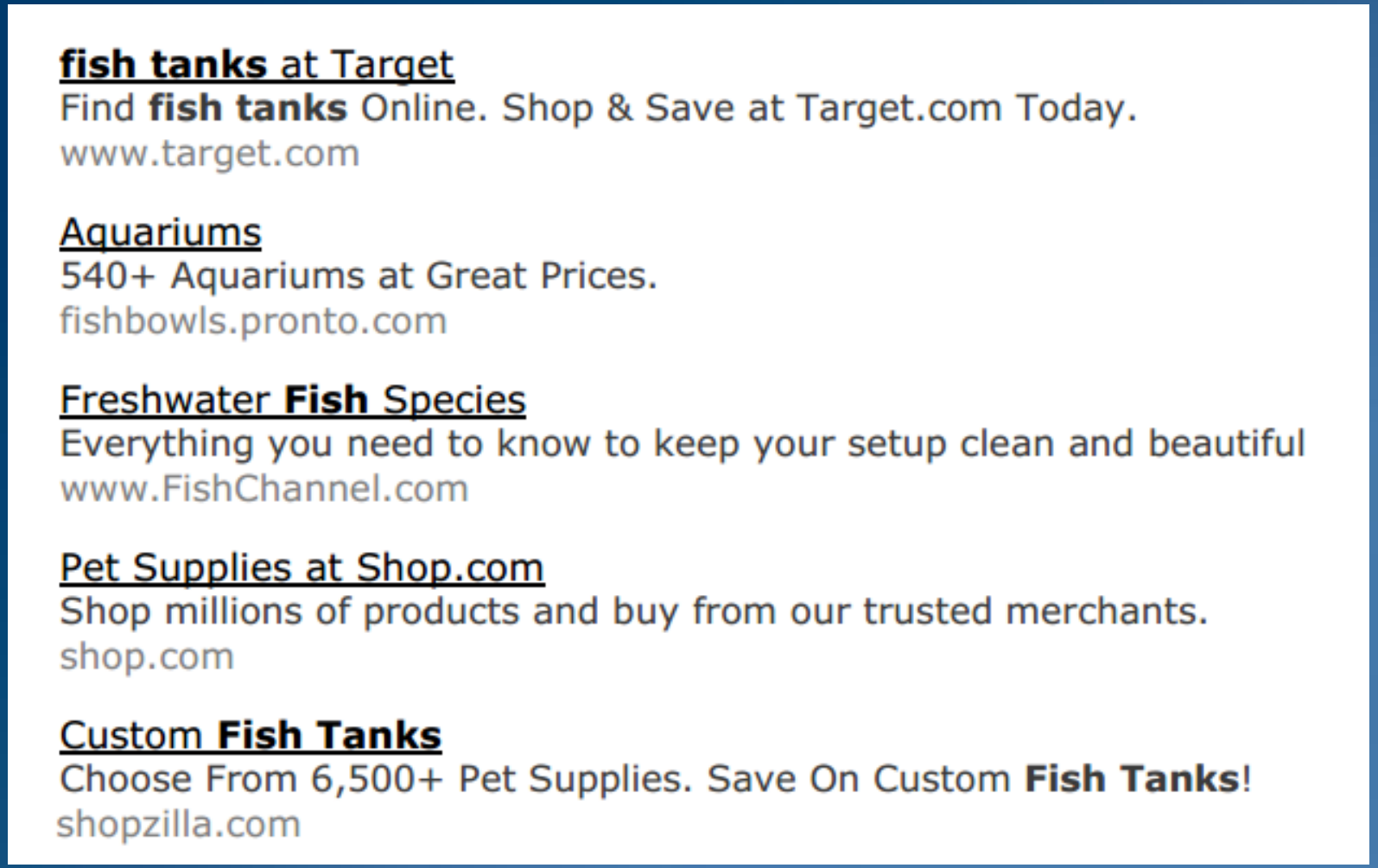
For example, the association between the phrases “fish tank” and “aquarium” may occur often in search sessions as users reformulate their original query to find more web pages. If this happens often enough relative to the frequency of these phrases, it will be considered significant.

Advertising and Search

The significant associations can be used as potential substitutions, so that, given an initial query, a ranked list of query reformulations can be generated, with the emphasis on generating queries that contain matches for advertising keywords.

Advertising and Search

The expansion technique consists of using the Web to expand either the query, the advertisement text, or both. A form of pseudo-relevance feedback is used where the advertisement text or keywords are used as a query for a web search, and expansion words are selected from the highest-ranking web pages.



Advertisements displayed by a search engine for the query “fish tanks”

Advertising and Search

As an example, the figure shows the list of advertisements generated by a search engine for the query “fish tanks”. Two of the advertisements are obvious matches, in that “fish tanks” occurs in the titles. Two of the others(the second and fourth) have no words in common with the query, although they are clearly relevant.

Advertising and Search

- Using the simple pseudo-relevance feedback technique described would produce both “aquarium” (frequency 10) and “acrylic” (frequency 7) as expansion terms based on the top 10 results. This would give advertisements containing “aquarium”, such as the second one, a higher relevance score in the selection process.

Advertising and Search

A more effective approach is to use a classifier based on machine learning techniques, A classifier uses a weighted combination of features to determine which words and phrases are significant.

Advertising and Search

Typical features include the frequency in the document, the number of documents in which the word or phrase occurs, functions of those frequencies, frequency of occurrence in the query log, location of the word or phrase in the document, and whether the word or phrase was capitalized or highlighted in some way. The most useful features are the document and query log frequency information.

Clustering the Results

If a user is interested in a particular aspect of a query topic, scanning through many pages on different aspects could be frustrating. This is the motivation for the use of clustering techniques on search results. Clustering groups documents that are similar in content and labels the clusters so they can be quickly scanned for relevance.

Clustering the Results

If a user is interested in a particular aspect of a query topic, scanning through many pages on different aspects could be frustrating. This is the motivation for the use of clustering techniques on search results. Clustering groups documents that are similar in content and labels the clusters so they can be quickly scanned for relevance.

Pictures (38)

Aquarium Fish (28)

Tropical Fish Aquarium (26)

Exporter (31)

Supplies (32)

Plants, Aquatic (18)

Fish Tank (15)

Breeding (16)

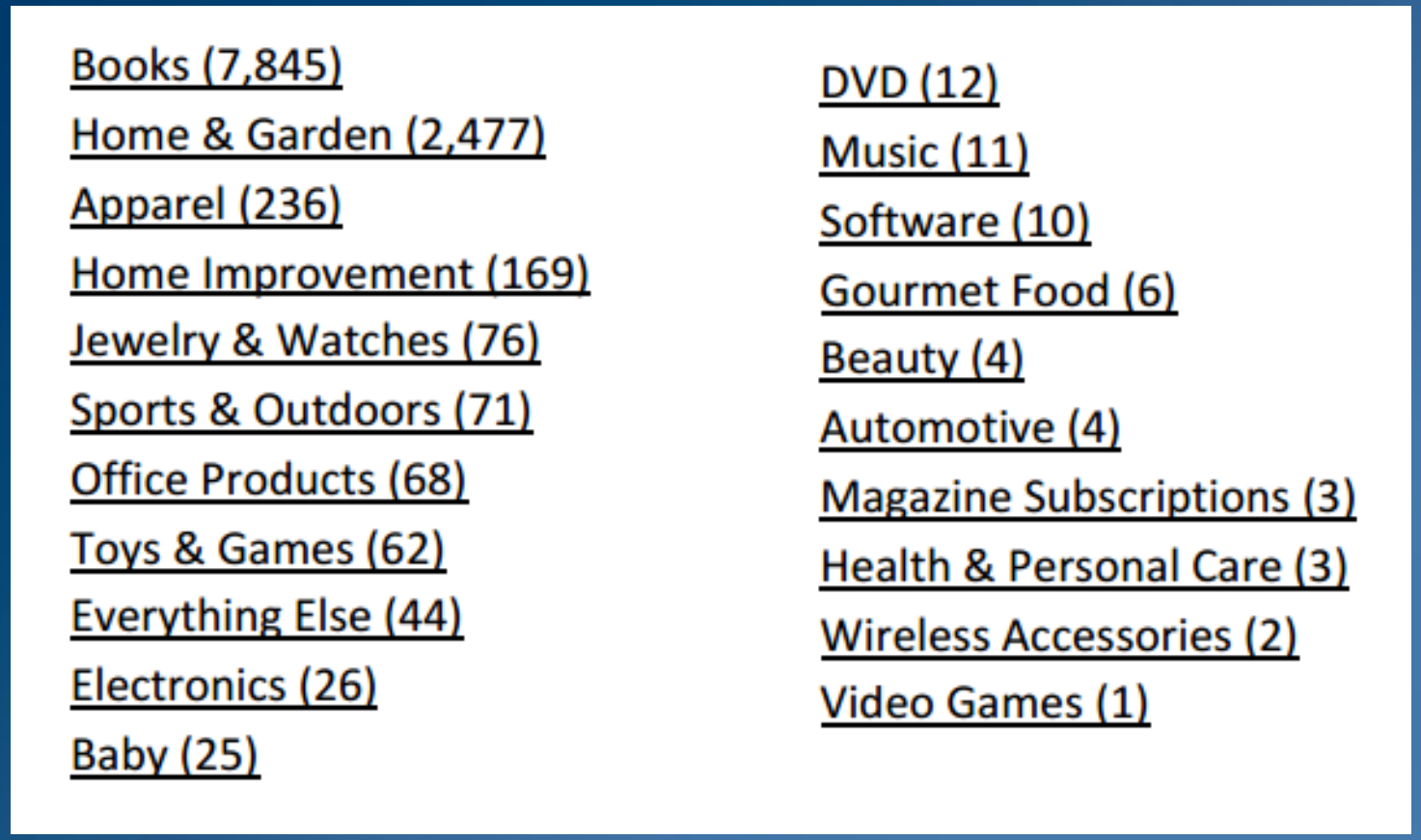
Marine Fish (16)

Aquaria (9)

Clusters formed by a search engine from top-ranked documents for the query “tropical fish”. Numbers in brackets are the number of documents in the cluster.

Clustering the Results

This list, where each cluster is described or labeled using a single word or phrase and includes a number indicating the size of the cluster, is displayed to the side of the usual search results. Users that are interested in one of these clusters can click on the cluster label to see a list of those documents, rather than scanning the ranked list to find documents related to that aspect of the query.



Categories returned for the query “tropical fish” in a popular online retailer

Home & Garden

Kitchen & Dining (149)
Furniture & Décor (1,776)
Pet Supplies (368)
Bedding & Bath (51)
Patio & Garden (22)
Art & Craft Supplies (12)
Home Appliances (2)
Vacuums, Cleaning & Storage (107)

Brand

<brand names>

Seller

<vendor names>

Discount

Up to 25% off (563)
25% - 50% off (472)
50% - 70% off (46)
70% off or more (46)

Price

$0-$24 (1,032)
$25-$49 (394)
$50-$99 (797)
$100-$199 (206)
$200-$499 (39)
$500-$999 (9)
$1000-$1999 (5)
$5000-$9999 (7)

Subcategories and facets for the "Home & Garden" category

Clustering the Results

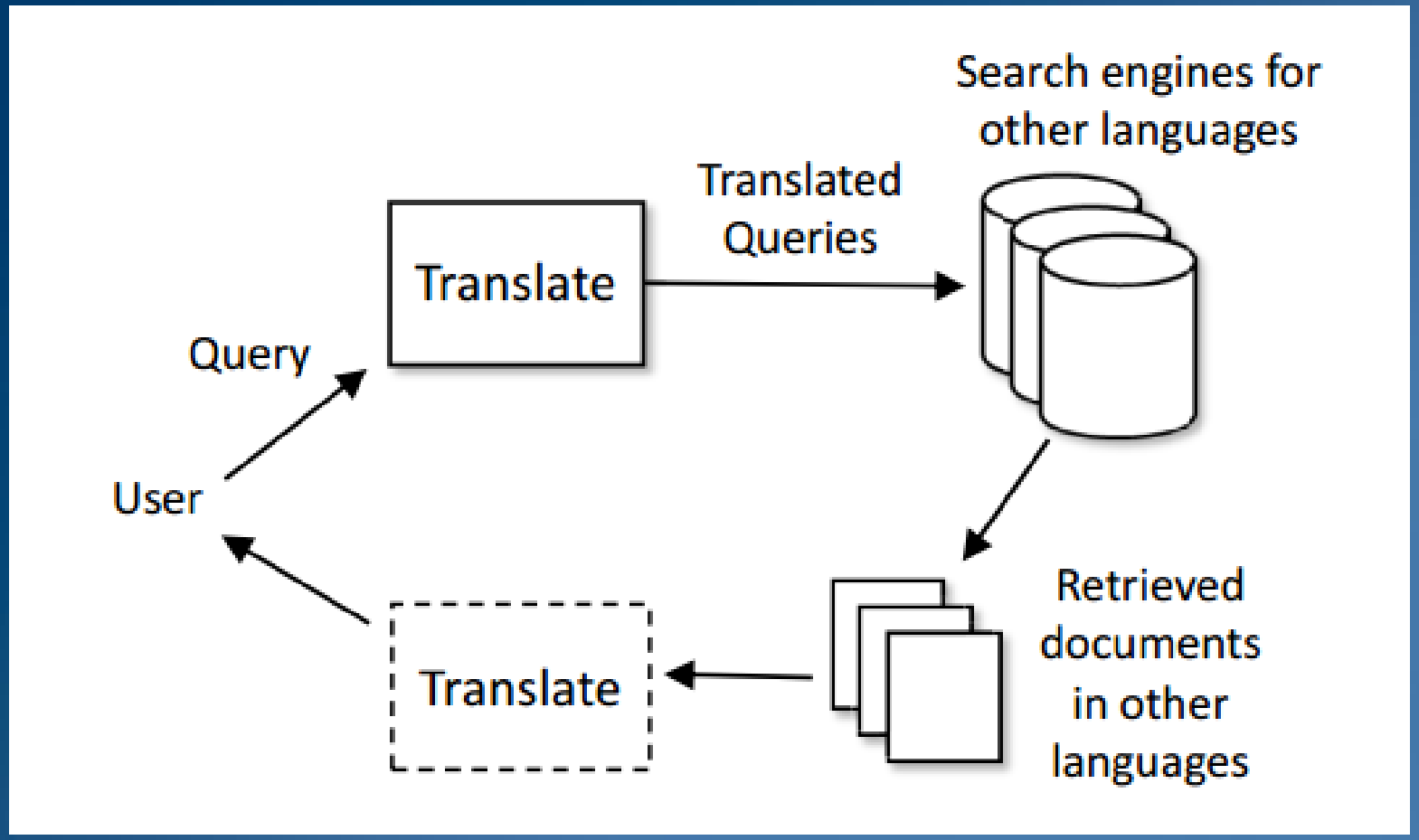
The numbers refer to the number of products in each category that match the query. These categories are displayed to the side of the search results, If the “Home & Garden” category is selected, the following figure shows that what is displayed is a list of subcategories, such as “pet supplies”, together with facets for this category, which include the brand name, supplier or vendor name, discount level, and price.

Clustering the Results

The numbers refer to the number of products in each category that match the query. These categories are displayed to the side of the search results, If the “Home & Garden” category is selected, the following figure shows that what is displayed is a list of subcategories, such as “pet supplies”, together with facets for this category, which include the brand name, supplier or vendor name, discount level, and price.

Cross-Language Search

By translating queries for one or more monolingual search engines covering different languages, it is possible to do cross-language search. Users typically will not be familiar with a wide range of languages, so a cross-language search system must do the query translation automatically.



Cross-language search

Cross-Language Search

The most obvious approach to automatic translation would be to use a large bilingual dictionary that contained the translation of a word in the source language(e.g., English)to the target language(e.g., French). Sentences would then be translated by looking up each word in the dictionary.

Cross-Language Search

The most effective and general methods for automatic translation are based on statistical machine translation models. When translating a document or a web page, in contrast to a query, not only is ambiguity a problem, but the translated sentences should also be grammatically correct. Words can change order, disappear, or become multiple words when a sentence is translated.

Cross-Language Search

Statistical translation models represent each of these changes with a probability. This means that the model describes the probability that a word is translated into another word, the probability that words change order, and the probability that words disappear or become multiple words. These probabilities are used to calculate the most likely translation for a sentence.

Cross-Language Search

The probabilities in statistical machine translation models are estimated primarily by using parallel corpora. These are collections of documents in one language together with the translations into one or more other languages. The corpora are obtained primarily from government organizations, news organizations, and by mining the Web, since there are hundreds of thousands of translated pages.