

# Coveo Enterprise Search 6

## Ranking – Technical Overview

Ranking is the art of sorting results according to their relevance to the submitted query. This article describes technical features of *Coveo Enterprise Search's* (CES) ranking engine. More specifically, it elaborates on the benefits of using automatic document summarization technologies in the ranking process.

### Objectives

There are three main objectives regarding the quality of the result list provided by a search engine following a user query:

1. A search engine should only return relevant information to a query.
2. The results should be ranked by relevance.
3. All relevant results should be returned by the search engine so that no important document is left out.

### Precision vs. Recall

Search engine performance is traditionally measured using two conflicting criteria: precision and recall. Precision is related to the first and second objectives, while recall is related to the third objective.

Actually, we can think of measuring precision as answering both the questions: *What proportion of documents returned by the search engine is relevant?* and *At which position is the document searched for situated in the results list?*. Similarly, seeking to increase precision corresponds to **trying to find relevant documents only and to rank them by relevance**.

On the other hand, measuring recall amounts to answering the question *What proportion of relevant documents are returned by the search engine?* and seeking to increase recall means **trying to find all relevant documents**.

Generally speaking, it is possible to improve precision (objective 1) and ranking (objective 2) simultaneously. However, this is not true for these factors versus recall (objective 3): increasing recall generally decreases precision and vice versa. Faced with this tradeoff, which element, precision or recall, should be favored? **We do believe that the focus should be put on precision**.

The following are arguments to support this statement:

- A large part of queries are actually performed by users that seek a particular document or Web site, something related to high precision.
- Even when many documents are potentially relevant to a search, most of the time, **the user will be satisfied with the first few relevant results**. Therefore, it is better to provide the results most likely to contain the required information (precision) than returning every result that include this information (recall).
- **Precision decreases faster than recall increases**; meaning that if a user really wants to see every single relevant result to his query, he will have to filter out many more irrelevant documents than the number of additional relevant documents he will have access to.

- **It is much more difficult to increase precision than recall.** For instance, if a user is not satisfied by his query's level of recall, he can formulate a new query that will retrieve more documents (using new keywords and the OR operators). This is not true for low precision systems: it will be tedious for the user to narrow down the results list if it is overwhelmed with irrelevant results.
- **Users mostly consult the first result page;** thus, increasing recall will be of no utility to a majority of users.
- Recall can easily be increased by means of a technique called *query expansion*, namely through the use of a thesaurus. **However, general thesauruses can hardly increase precision.** Consequently, we believe that the focus must be kept on precision as, if required, recall can be enhanced by a thesaurus that is independently developed and operated by the organization's content manager, yet the same is unlikely for precision.

CES is designed and optimized to fulfill, in a priority order, the three objectives stated at the beginning of this section.

#### ▶ Successful Queries

For CES, a successful query has the following characteristics:

1. The first result returned by the search engine is the most relevant to the user's query;
2. The first result page mostly consists of highly relevant results;
3. Even under stress conditions, a query should be answered right away<sup>1</sup>. Note that satisfaction of objectives 1) and 2) should not be done to the detriment of speed.

#### Summarization and Key-Concept Extraction

CES's ranking feature is partly based on summarization and key-concept extraction techniques. Traditional search engines merely match query-words with document-words during the retrieval process. Then, simple techniques are used to rank results. For instance, given the query *radio systems for car*, the most important words are *radio* and *car*. Actually, *for* is not very informative (it is found in almost every document) and *systems* is rather general. Thus, it appears natural to give more importance in the ranking process to *radio* and *car* than to *for*. For this reason, a common weighting measure has been commonly used: TFIDF (or variants, such as Okapi BM25). The TFIDF formula is:

$$weight_{t,d} = \begin{cases} \log(tf_{t,d} + 1) \log \frac{n}{x_t} & \text{if } tf_{t,d} \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

$tf_{t,d}$  is the frequency of word  $t$  in document  $d$ ,  $n$  is the number of documents in the text collection and  $x_t$  is the number of documents in which the word  $t$  occurs. Accordingly, the more often a query-word appears in the document, the more important this word is for that document. On the other hand, the more often the word appears in the text collection, the less informative this word becomes.

TFIDF, or term word weighting variants, have proved to be generally useful to assess word importance. But formulas such as TFIDF have several drawbacks:

<sup>1</sup> In an index containing 1,000,000 documents, standard two-word queries should be answered in less than one second when the query load is no more than 14,000 queries an hour. Under normal usage, most queries should be answered in the 50ms to 250ms range.

- **Words are weighted regardless of their position in the document.** A word displayed for the first time in the last paragraph of a 1000-page document gets the same weight as if it is in the first line of another document. Moreover, words are sometimes displayed in places where they are not related to the actual textual content of the document. Consider, for instance, the text related to advertisements on Web pages or non-textual data in Excel worksheets.
- The **context of the word is not considered.** Take for instance, the query *car radio*. Formulas such as TFIDF give the same score to the sentences *New car radio model available* and *Car makers use the radio a lot to advertise*.
- **Word counts are based on exact word match** only. TFIDF counts *Mary* once in ***Mary** has a conference call appointment at 10h AM, thus **she** will be late at the meeting* even if *she* refers to *Mary*.
- **Word counts do not measure rhetorical subtleties.** Bulleted lists in Word documents often contain worthy information. Or, for instance, *Radio XYZ* is more important in *Thus, all owners of radio XYZ should return it for a fix* than in *If the radio noise problem originates in the car's electrical system you may have to install a noise filter on your radio (radio XYZ models already include one)*. Words in conclusive remarks are more informative than words in email signature quotes. These are all examples of rhetorical subtleties that are not captured by simple formulas.

Coveo's text summarization technologies pinpoint the key concepts and extract the most relevant sentences, thus overcoming TFIDF-like measure shortcomings. As a result, the resulting ranking is much better than traditional search engines.

### Customizable Multi-Criteria Ranking

Key-concept and sentence extraction is not the sole criterion by which pages are ranked. In fact, there are many factors that contribute to the final score. Each ranking factor can be adjusted (if required) by the content manager to better suit each corporation specific needs.

The following lists some of these parameters:

- ▶ **Weights that depend on the query terms:**
  - **Term proximity.** Query words in documents in the same context (that is, near each other) get a higher score than if they were not.
  - **Term frequency.** TFIDF-like measure. Frequent words in document, but rare in the whole text collection obtain the highest weight.
  - **Term in concepts.** If a query word matches a key concept extracted by Coveo's key-concept extraction technology.
  - **Term in summary.** If a query word is found in a key sentence extracted by Coveo's summarization technology.
  - **Term matches original word.** If a query term that matches a word in a document before it was modified for the index (accents, casing and stemming).
  - **Term in title.** If a query word matches words in the title.
  - **Term in address.** If a query word matches words in the document address (URI) (for example, *models* is matched with *q:\products\models\XYZ\specs\1998*).

- **Term in home page address.** If a query word matches words in the home page document of a Web site. For example, the home page of *Coveo* ([www.coveo.com](http://www.coveo.com)) for the query *Coveo*.
- **Term is highlighted.** Highlighted words in documents (**bolded, underlined, bigger font**) get a boost.
- **Term in domain name.** If a query word matches the domain name of a Web home page. For instance, *krystal* matches *audio.krystal.com* (but not *audio.krystal.com/support*).
- **Term in folder name.** If a query word matches the last folder name where the document is located. For instance, in the query *radio model XYZ*, *XYZ* matches documents in *q:\doc\specs\models\XYZ\*, but not *q:\doc\specs\models\XYZ\specs\1998\* (however, *XYZ* gets points for the address match above).
- ▶ **Weights that are document-dependant only, meaning regardless of the query terms sent to the search engine:**
  - **Document modified recently.** More recent (or recently modified) documents get more points.
  - **Document quality evaluation.** Pages located near the root of the address path get more points. For instance, *audio.krystal.com* gets more points than *audio.krystal.com/models/XYZ/support/old\_site/faq.html*.
  - **Source rating.** Some text sources contain more valuable documents than others. The content manager can define a rating modifier to take this into account during the ranking process.
  - **Document in user language.** Points are given to the document if its language matches the user language preferences.
  - **Collaborative Rating.** Users's personal rating of documents (clicked stars) influences results of users. Visited documents also modify a document rating over time.
  - **Custom ranking weigh.** Document rating can be altered during the indexing process by an administrator's script.

### Balancing Scores Between Documents of Different Source Types

In addition to the set of ranking factors described above, certain weights are modified according to the type of each document to maintain balance between scores. This ensures that all documents (emails, documents on a file system or Web pages) have similar chances of ranking first. Some examples of score inconsistencies that impel this normalization process follow:

- The modification date weight is set much higher for files and e-mails than Web pages: thus, files are likely to receive more points;
- Web pages and Microsoft Office™ documents can have font modifier scores (for instance, bold) while plain text emails and text files cannot.