Hewlett Packard Enterprise

HP Propel

Software version 2.10

HP Propel Search with IDOL

A white paper

Contents

Documentation release date: December 2015

Software release date: December 2015

Legal Notices

Warranty

The only warranties for Hewlett Packard Enterprise products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. HPE shall not be liable for technical or editorial errors or omissions contained herein. The information contained herein is subject to change without notice.

Restricted Rights Legend

Confidential computer software. Valid license from Hewlett Packard Enterprise required for possession, use or copying. Consistent with FAR 12.211 and 12.212, Commercial Computer Software, Computer Software Documentation, and Technical Data for Commercial Items are licensed to the U.S. Government under vendor's standard commercial license.

Copyright Notice

Trademark Notices

Adobe® is a trademark of Adobe Systems Incorporated.

Microsoft® and Windows® are U.S. registered trademarks of Microsoft Corporation.

Oracle and Java are registered trademarks of Oracle and/or its affiliates.

UNIX® is a registered trademark of The Open Group.

RED HAT READY™ Logo and RED HAT CERTIFIED PARTNER™ Logo are trademarks of Red Hat, Inc.

The OpenStack word mark and the Square O Design, together or apart, are trademarks or registered trademarks of OpenStack Foundation in the United States and other countries, and are used with the OpenStack Foundation's permission.

Documentation Updates

The title page of this document contains the following identifying information:

- Software Version number, which indicates the software version.
- Document Release Date, which changes each time the document is updated.
- Software Release Date, which indicates the release date of this version of the software.

To check for recent updates or to verify that you are using the most recent edition of a document, go to the following URL and sign-in or register: https://softwaresupport.hp.com/

Use the Search function at the top of the page to find documentation, whitepapers, and other information sources. To learn more about using the customer support site, go to: https://softwaresupport.hp.com/documents/10180/14684/HP_Software_Customer_Support_Handbook/

You will also receive updated or new editions if you subscribe to the appropriate product support service. Contact your Hewlett Packard Enterprise sales representative for details.

Support

Visit the Hewlett Packard Enterprise Software Support Online web site at https://softwaresupport.hp.com/

This web site provides contact information and details about the products, services, and support that HPE Software offers.

HPE Software online support provides customer self-solve capabilities. It provides a fast and efficient way to access interactive technical support tools needed to manage your business. As a valued support customer, you can benefit by using the support web site to:

- Search for knowledge documents of interest
- Submit and track support cases and enhancement requests
- Download software patches
- Manage support contracts
- Look up HPE support contacts
- Review information about available services
- Enter into discussions with other software customers

- Research and register for software training

To learn more about using the customer support site, go to:
https://softwaresupport.hp.com/documents/10180/14684/HP_Software_Customer_Support_Handbook/

Overview

Autonomy Intelligent Data Operating Layer (IDOL) Server integrates unstructured, semi-structured, and structured information from multiple repositories through an understanding of the content. It delivers a real-time environment to automate operations across applications and content, removing all the manual processes involved in getting information to the right people at the right time.

HP Propel uses IDOL's conceptual search (basic query) to search for the search terms. In a basic query, you provide keywords or sentences, and IDOL Server retrieves the documents that match it. These queries are the easiest type of search for end users, because they do not require any special syntax or training. IDOL will process the search term as follows. Note that these processes are described in more detail later in this document.

- Removes the stop words from the query text.
- Stems each word to a linguistic root.
- Applies transliteration.
- Expands synonyms.
- Matches the text against the index fields of the documents it contains.
- Returns documents that contain any or all of your search terms, in any matching form.

The relevance of a document depends how many tokenized term are matched as well as the number of occurrences. Since HP Propel is using advanced search, it also depends on how close the matched terms are. Search term matches are determined by many factors. Stemming, transliteration, character normalization, synonyms are enabled out-of-the-box (OOTB) in Propel to enhance the search experience. Stemming matches a document when it contain a word with the same stem. Transliteration and character normalization aid searching across a language, especially when a character has more than one form. Synonyms is a user defined list of words to generalize the search term., meaning that the user does not have to know the exact word used in a document but any one of its synonyms to find matches.

HP Propel also supports exact phrase search, specified by enclosing the search term with double quotation marks. Instead of finding the exact phrase in the document, IDOL (in fact, search engines in general) tries to help user find the relevant documents without having to know the exact wordings in the document. Enclosing a search term in double quotes, overrides this default behavior and forces IDOL to search only for the exact search term.

HP Propel indexes multiple fields in service catalog items, support catalog items, and knowledge articles, including title, description, content, attachments and form data. Note that Propel indexes more fields than what is shown in the search results page, so it is possible to have results that don't appear as an obvious match because, for example, the search term matched content in an attachment and not in the title.

HP Propel doesn't boost the weight of any of the indexed fields, so a match in the title, for example, has the same weighting as a match in the description. If a document mentions a search term twice in the description, it will be considered as more relevant than another document that only mentioned the search team once in the title.

Conceptual Search

IDOL Server accepts a piece of content (a sentence, paragraph or page of text, the body of an e-mail, a record containing human-readable information, or the derived contextual information of an audio or speech snippet) or reference (identifier) as input and returns references to conceptually related documents ranked by relevance, or contextual distance. IDOL Server uses this process to generate automatic hyperlinks between pieces of content.

Conceptual search includes keyword and natural language searches, which allow you to specify terms that you want to find in your document set. IDOL applies some processing to the search terms so that it can find related terms:

Language type

HP Propel has turned off automatic language detection in IDOL; each document is indexed with a language type defined in `AutonomyIDOLServer.cfg`. Note that the configuration file defines a number of language types that IDOL recognizes, and one language type will be specified for each document. Propel also uses the language type to search only the document that matches the language settings on the user's browser. Consider the following situations:

- If an item is available in three languages and a search term is found in all three, the language setting in the user's browser will determine which version of the document will be considered a match to the search term.
- If the search term matches only one of the documents (probably because the term was translated, and therefore not identical in the other versions of the document), the document that matches the term will be indicated as a match regardless of the browser's language setting. For example, if the English word *server* is translated as *serveur* in the French version of the document, and the user searches on the term *serve*r, only the English version of the document will be considered a match even if the user's browser is configured with French as the selected language.

Note that the above content is only for the languages Propel supports.

Tokenization

IDOL Server stores document text as a series of tokens. Generally, a token is a word, but it can also include other strings of characters, for example, a phone number or e-mail address. During the indexing process, IDOL Server converts the text into tokens for matching, and stores it in Index fields.

IDOL Server processes characters according to their common use, and uses them to define tokens in text.

In IDOL Server, the following three types of characters define tokens and the breaks between tokens:

CHARACTER TYPE	DESCRIPTION	WHEN INDEXED INTO IDOL SEARCH
Text character	Letters and numbers, including logograph characters from Asian writing systems.	Does not change
Separator character	Characters that separate two words, such as spaces, tabs, and line breaks.	Become spaces, which mark the break between one token and the next.
Non-separator character	Other characters, such as punctuation.	Deleted. If text is separated only by non-separator characters, the text becomes a single token.

Consider the following e-mail address: joe.smith@example.com. In the default IDOL Server configuration:

- The at symbol (@) is a separator character.
- The period (.) is a non-separator character.

When IDOL Server processes the e-mail address, it produces two tokens: JOESMITH and EXAMPLECOM, which you can search for to return this document as a match. However, if you search for JOE or EXAMPLE, this document is not considered a match.

The default list of separator characters can be configured via `TangibleCharacters`. `HyphenChars`, `AugmentSeparators` and `NumberPunctuation` will also change the way IDOL indexes data. HP Propel use the default settings for all these parameters.

Stemming

Stemming is the process of reducing a word to its linguistic root. The purpose of this reduction is to find a base term, so that a search can be expanded to include all forms of the term. For example, you generally want a search for the word *elections* to match a document that contains the word *election*. As long as the two terms stem to the same form, both return in the search. During indexing, IDOL Server stems each term, and stores the stem as well as the unstemmed term. During querying, IDOL Server stems the query term, and matches it against the stored stems in the index.

The user can search for unstemmed (exact) terms by enclosing the search term in double quotation marks ("). For example:

- A search for *election* matches both *election* and *elections*.
- A search for *"elections"* matches only the exact form, *elections*.

Stemming can be customized using a stemming file. Propel does not customize stemming OOTB.

Stop List

Stop lists are sets of words (called *stop words*) that are ignored when processing both index fields at index time, and subsequently as part of queries. The primary reason to use stop lists is to reduce index size and make queries faster by ignoring words that add little or no meaning to the retrieval process.

HP Propel uses the OOTB stop list for all languages.

Character Normalization

In some cases, there is more than one way to represent a character. For example:

- The Roman alphabet has uppercase and lowercase forms of all letters.
- The Japanese katakana script can have full width or half width characters.

IDOL Server uses canonicalization to ensure that it treats all character forms equally. It automatically converts to an internationally recognized canonical form.

Transliteration

Transliteration is a sort of character normalization in that it aims to map sets of characters to a standard form so that a search for different forms match documents containing any of those forms. An important example is the removal of accents from letters so that a search for *cafe* matches documents containing *café*, and vice versa. Similarly, the German letter ß is transliterated to *ss*.

HP Propel has transliteration turned on (this also controls character normalization).

Decomposition

Decomposition breaks single compounded words into multiple parts to enable searches to match parts of the word. Decomposition rules are particularly useful in languages such as German and Hungarian. New decomposition rules should contain compound words that are tokenized, transliterated and stemmed, e.g., mousetrap → mouse trap

HP Propel doesn't provide any OOB decomposition rules, nor any command line tools to manage decomposition rules.

Boolean and Bracketed Boolean Search

IDOL Server accepts simple or complex Boolean and bracketed Boolean expressions. Form search terms with Boolean expressions using the following set of Boolean and proximity operators:

AND	XOR/EOR	WNEAR
NOT	NEAR	BEFORE

OR	DNEAR	AFTER

Proximity search operators WNEAR, NEAR and DNEAR, consider the number of words between two specified words to determine their proximity to one another. If two words are adjacent to each other, their proximity is 1. If one word separates them, their proximity is 2, and so on.

NEAR returns only documents in which the second search term is within *N* words of the first term. DNEAR – directed NEAR – returns only documents in which the second term is within *N* words of the first term, in the specified order.

WNEAR - Weighted NEAR (with OR operation) - is the default proximity search operator as HP Propel has advanced search enabled. This proximity operator returns documents that contain either of the two terms. It promotes relevance when the terms are *N* or fewer words apart (closer together implies higher relevance). If you do not specify *N*, WNEAR defaults to 5. For example, query `action=Query&Text=red WNEAR7 green`, returns documents that contain either red or green. It gives extra relevance to documents in which red and green appear seven or fewer words apart in a piece of text. This weight increases as the terms get closer to each other. Documents in which the terms occur more than seven words apart, or in which only one term occurs, return with normal relevance.

Wildcard Search

Use wildcards in IDOL Server to find terms that are similar to the search term. In wildcard searches, a question mark (?) represents a single missing character, and an asterisk (*) represents any number of missing characters.

When IDOL Server receives a wildcard query, it removes stop words and stems the whole terms as usual. It then expands each term that contains a wildcard to a list of terms that IDOL Server contains that match the wildcard.

When it expands a wildcard, IDOL Server finds all the exact (unstemmed) terms in the index that match the wildcard. If the wildcard is not very specific, or if you have a large number of terms in the unstemmed index, this process can take a long time.

After it expands the wildcard to all the unstemmed terms, IDOL Server retrieves all the documents that contain these terms, in addition to any other terms in the query.

Advanced Search

`AdvancedSearch` enables exact stem matching as part of a phrase query. `AdvancedCaseSearch` enables exact case matching. To use this feature, you prefix terms with a tilde (~), and IDOL treats the terms as case-sensitive. `AdvancedPlus` mode enables the PARAGRAPH and SENTENCE operators, and indexes terms with both an explicit weight and position (so that position and weight information is available to subsequent queries).

For example, consider the following text: *Searching terms with weights[90]*

CONFIGURATION	QUERY	RETURNS?
AdvancedSearch=FALSE	"search terms"	Y
AdvancedSearch=TRUE	"search terms"	N
AdvancedSearch=TRUE	"searching terms"	Y
AdvancedSearch=TRUE	~searching	Y

AdvancedCaseSearch=TRUE	~searching	N
AdvancedCaseSearch=TRUE	~Searching	Y
AdvancedCaseSearch=TRUE	"terms with weights"	N
AdvancedPlus=TRUE	"terms with weights"	Y

HP Propel sets AdvancedSearch=TURE, i.e., IDOL won't do stemming for terms enclosed in double quotes.

The Order of Language Process

During indexing,

1. Automatic language detection (disabled in Propel)
2. Sentence breaking and NGram tokenization (primarily for Asian languages, Propel using OOB settings)
3. Tokenization
4. Normalization and transliteration
5. Stop word detection
6. Stemming
7. Decomposition

During query,

1. Query Manipulation (QMS) including synonyms.
2. Synonym Content Engines (not used in Propel)
3. Decomposition

It is important to understand the order which IDOL processes the search query, especially when customizing IDOL. HP Propel uses OOTB settings for most features. See individual feature discussion previously in this document.

Relevancy Calculation

The core relevancy calculation takes into account the APCM weight (statistical weight) assigned to each of the matched terms, and the number of times those terms occur in the result document.

Matching more of the terms generally results in a higher ranking than matching a few of the terms multiple times. Additional factors such as capitalization, stemming, and proximity weighting also adjust the weighting (but again, normally less than matching an additional term).

The advanced search modes (AdvancedSearch, AdvancedCaseSearch, or AdvancedPlus) change the default query operator between terms to use WNEAR where no query operator is specified (rather than OR in the normal search mode). In this case, documents matching the query terms in proximity to each other receive a weighting boost, with a larger boost if consecutive query terms appear as a phrase in a document. This process is distinct from an exact phrase search (quoting the query terms as a phrase), which returns documents only if they match the exact phrase specified.

Field Weights

Fields can be assigned differing weights to boost one field over others, e.g., boosting the title field over content field.

The most common reason to change occurrence weighting is to allow or prevent term counting from dominating the result weighting. By default, IDOL relevance calculations balance the occurrence counting with other factors.

HP Propel does not have boosting for any index field, i.e., matches in all index fields are weighted equally. Propel uses BIAS to boost the relevancy to retrieve the matching document in the correct language.

Parametric Search

Parametric search (also known as faceted search) allows you to find documents that have a particular value or range of values in a particular field. This type of search is useful when you want to be able to find a set of documents with particular properties.

For example, if you have a database of products, you might want users to be able to search for *toaster* and then restrict the results by brand, model, color, power usage, and energy efficiency. With parametric search, you can return a list of the brands, models, and so on that exist for all products that match the search *toaster*. Users can then restrict the search by a particular value.

Parametric Search Process

To use parametric search, configure all fields you want to use as parametric fields. Then use IDOL Server actions to:

- Return a list of all parametric fields in IDOL Server.
- Return a list of all values that occur in a particular field for all documents.
- Return a list of all values that occur in a particular field for documents that match a particular search.
- Return a count of the number of documents that contain a particular value in a particular field.

After you find the value you want to filter by, you send a Field Search to IDOL Server to retrieve documents that have this value in this field.

HP Propel uses parametric search to generate the side navigation bar.

Synonym Search

You can add a list of synonyms for search terms to IDOL Server. When you search for a term, IDOL Server automatically includes all the synonyms for that term. Synonyms allow you to broaden a search and return documents on a subject that do not contain the exact word. This increases the power of natural language queries.

HP Propel uses the QMS to perform synonym search. QMS acts as a proxy between the user interface and the data Content server. It communicates with both the data Content server, and an AgentStore component that holds the synonym rules. Using QMS is scalable and dynamic. It also reduces the amount of knowledge that the user interface requires of the server implementation. You can also use this method to substitute phrases in or out of the query text.

Typeahead Search

HP Propel's typeahead feature uses IDOL TermExpand API to expand the search term, Propel will enclose the search term with *, e.g., for search *text, ppl*, Propel will call IDOL's term expand with *ppl* which will return matching term from all databases in IDOL.

Learn more at

hpe.com/software/propel

Sign up for updates

December, 2015

Hewlett Packard Enterprise