Sintassi ricerche

Sintassi di ricerca linguaggio nativo eXtraWay

Sintassi di ricerca eXtraWay

La presente documentazione non ha lo scopo di chiarire le nozioni più elementari inerenti la sintassi adottata del server eXtraWay (ed ereditata direttamente dal server HighWay) per cui trattera questi aspetti molto rapidamente.

Sintassi di ricerca: Indicazioni di base

In primissimo luogo la ricerca può essere Globale o Mirata. Si considera Ricerca Globale una ricerca che non fa riferimento ad un particolare canale di ricerca e quindi che viene estesa automaticamente a tutti i canali di ricerca disponibili. Mentre in HighWay ciò aveva senso compiuto, in eXtraWay si tende a non fare solitamente uso di questo tipo di ricerca privilegiando la ricerca per il cosiddetto campo chiocciola che, in buona sostanza, consiste nell'estendere la ricerca dei termini impostati in selezione ad un sottoinsieme definito dei canali di ricerca disponibili.

Concentriamo comunque la nostra attenzione sulla ricerca classica ovvero quella nella quale si compie regolarmente l'indicazione dei canali desiderati e dei termini di ricerca nelle loro diverse combinazioni.

In primo luogo vediamo come si esprime la ricerca elementare:

```
<Nome Canale esteso o alias Canale con eventuali modificatori> <Operatore di Ugualizanza/Disuguaglianza>
<Termini di ricerca>
```

L'identificazione del canale.

Il Nome del Canale va racchiuso **obbligatoriamente** tra parentesi quadre ed il suo aspetto sarà rappresentato dal nome della componente principale di ricerca (XML per il contenuto dell'unità documentaria e UD per i dati di servizio di tale unità), una virgola ed il percorso della chiave XML corrispondente. Quanto detto può essere globalmente sostituito da un alias di ricerca da configurare appositamente nel file di configurazione d'archivio (nomearchivio.conf.xml)

Il nome del canale può a sua volta essere seguito da uno o più modificatori. Essi causano l'applicazione al singolo canale di una qualche caratteristica di ricerca. L'elenco e la spiegazione dei possibili modificatori è in appendice.

- **Nota:** Si noti che questi due metodi di ricerca, con il percorso XML e con il serach alias, conducono a due diversi comportamenti in caso di errata espressione.

  Se si esprime, in frase di ricerca, un search alias che non è stato opportunamente configurato nel file di configurazione d'archivio, il server torna una condizione d'errore (errore `820`, ovvero Campo non Chiave) per indicare l'errata espressione.

  Se invece si esprime il percorso di chiave XML o UD come nome di campo e ad esso non corrisponde un canale di ricerca riconosciuto il server non da segnalazioni di alcun tipo e considera la ricerca come se la condizione espressa su tale canale fosse sempre **falsa**.

  Questo dipende dal fatto che nel file di configurazione è obbligatorio indicare solo i canali per i quali non si applicano le configurazioni di default e quindi nulla vieta che un canale di ricerca, previsto o prevedibile, al quale si applichino le impostazioni di chiave di default e quindi non citato nel file di configurazione, non sia stato ancora rilevato in alcun documento. In questo scenario, un'applicazione che poggi su un simile archivio e faccia uso di quel canale compie una selezione lecita e sarebbe improprio che il server tornasse un errore in questo frangente. Malauguratamente è impossibile distinguere il percorso sbagliato dal percorso semplicemente sottinteso e non ancora incontrato.

Il nome del canale, in alcuni casi, può essere anche un valore speciale ovvero [?SEL], [?COL] o [?NDOC]. Il primo serve ad indicare una precedente selezione come fattore di restrizione (se in AND) o ampliamento (se in OR) combinato con gli altri termini di ricerca. Il valore da indicare come termine di ricerca sarà l'ID della selezione stessa. Il nome della selezione può essere espresso senza alcun percorso e senza l'estensione di default

.tmp.

Il secondo ha lo stesso scopo e lo stesso funzionamento ma fa riferimento alle raccolte ed il valore da impostare sarà appunto il nome della raccolta da utilizzare (essendo una raccolta una sorta di file di selezione non volatile).

Il nome della raccolta può essere preceduto dal carattere '+' ad indicare che la raccolta in esame non è una raccolta pubblica bensì una raccolta privata dell'operatore che sta compiendo la selezione.

- **Nota:** Il nome della raccolta può essere evinto dall'elenco che l'interfaccia di consultazione deve mettere a disposizione ma si ricorda che ogni carattere di interpunzione ed ogni spazio viene convertito nel carattere '_'.

Il terzo identifica il numero fisico del documento. Il server eXtraWay, come si sa, deriva direttamente dal server HighWay e per lungo tempo ne ha emulato il comportamento. Nello svolgere questo compiuto ha reso disponibile per lungo tempo un canale di ricerca denominato `NRECORD` che voleva corrispondere al valor logico del documento. In eXtraWay questo valor logico non ha senso perché è troppo dipendente dalla struttura dell'XML che compone il documento. Il ruolo di questo canale è stato quindi cambiato nel selettore per numero fisico. Ciò nonostante, in applicazioni come xdocwaydoc, esiste un `search_alias` avente lo stesso identico nome: nrecord.

In questo modo il server non sa bene a quale dei due canali far riferimento ma, per coerenza, lascia che a prevalere sia il `search_alias` dell'applicazione.

Per consentire, nonostante questo, la ricerca per documento fisico, essa può avvalersi del campo denominato “?NDOC”.

In sintesi, se non esiste un `search_alias` denominato `nrecord`, ricercare per…

[?NDOC]=…

…equivale in tutto e per tutto a ricercare per…

[NRECORD]=…

SearchSint_BaseOperator Operatori di Uguaglianza e Disuguaglianza

Una volta chiusa la parentesi quadra per isolare il nome del Canale essa **deve** essere seguita da un operatore di uguaglianza o disuguaglianza come i seguenti:

- = Uguale
- <> Diverso
- >= Maggiore o Uguale
- > Maggiore
- Minore o Uguale
- < Minore

Gli operatori logici

A seguito dell'operatore di Uguaglianza o Disuguaglianza si devono indicare uno o più termini (eventualmente separato da operatori logici o combinati in range o elenchi). La rappresentazione di questi termini **è alla base del comportamento in ricerca** del Server eXtraWay.

Come detto i termini di ricerca possono essere separati, ma sarebbe più opportuno dire congiunti, da operatori logici quali:

- E Operatore AND, forma italiana
- AND Operatore AND, forma internazionale
- ADJ Opertore di Adiacenza.
- XOR Operatore eXclusive OR, forma internazionale
- O Operatore OR, forma italiana
- OR Operatore OR, forma internazionale

Mentre i più comuni operatori hanno un senso ben chiaro, l'operatore eXclusive OR comporta che l'Unità Informativa venga eletta se esiste un termine e non l'altro o vice versa ma non entrambe.

Gli opertori logici vengono valutati nell'ordine in cui sono stati elecati. Di fatto gli operatori E, AND e ADJ hanno pari valenza, poi viene valutato l'operatore XOR ed in fine, con pari valenza, gli operatori O ed OR.

Le estensioni di ricerca

Ogni ricerca che può essere effettuata prevede l'uso di estensioni di diverso tipo. In questo paragrafo si darà evidenza di ciascuna di esse.

In realtà alcune di quelle che seguono sono modalità di restrizione, più che di estensione, o comunque di supporto alla ricerca.

- **Uso di Wild Cards**

La più comune delle estensioni è senza dubbio l'estensione con l'utilizzo di Wild Cards. Gli operatori ne apprendono l'utilizzo con semplicità ed è senza dubbio la forma di estensione più frequente e diffusa. Chiunque abbia dimestichezza con l'uso di regular expression sa che risultati possa dare questo approccio.

Le Wild Cards adottate da eXtraWay sono l'asterisco '*' ed il punto interrogativo '?'. Al primo viene associato il concetto di “0, 1 o più caratteri qualsiasi” mentre al secondo il concetto di “un qualsiasi carattere”.

Possono essere dislocati in qualunque punto della chiave oggetto di selezione, ripetuti in più punti, ed il solo accorgimento riguarda l'uso dell'asterisco come primo carattere della chiave espressa.

Se utilizzato come primo carattere della chiave, l'asterisco può portare ad un'analisi del vocabolario dei termini indicizzati pressoché totale con evidenti effetti sulle prestazioni in ricerca. E' quindi facoltà di chi amministra il Data Base inibire questo comportamento richiedendo che l'uso delle Wild Cards sia limitato e che si imponga la presenza di un certo numero di caratteri reali in testa alla chiave espressa in ricerca.

Se, ad esempio, si impongono almeno 2 caratteri validi, una ricerca per…

```
[testo]=do*to
```

…sarà considerata valida mentre una ricerca per…

```
[testo]=p*collo
```

…verrà rifiutata come sintatticamente non accettabile.

- **Estensione per genere e numero**

In ordine di utilizzo l'estensione per genere e numero (vale a dire maschile, femminile, singolare e plurale) è indubbiamente paritetica all'utilizzo di Wild Cards.

Algoritmi basati sulla linga italiana, che possono essere rapidamente e semplicemente estesi ad altre lingue, in particolare all'inglese, consentono di ridurre un termine oggetto di ricerca alla sua forma abbreviata e da essa risalire alle 4 forme teoriche: maschile e femminile per la forma singolare e plurale. Va da se che questo tipo di estensione può produrre termini inesatti (es. la parola `casa` produce correttamente `case` ma conduce anche a `caso` e `casi` che esprimono in realtà concetti diversi) ma offre un comportamento più corretto dell'uso di una Wild Card in coda.

Si potrebbe infatti supporre che l'estensione per genere e numero consista nell'abbreviazione di un termine e nell'applicazione ad esso della Wild Card asterisco o punto interrogativo. Questo non porterebbe allo stesso risultato (es. il termine `pesca` non condurrebbe a pesche con l'uso di '?' ma imponendo uso di '*' si otterrebbero ben altri termini come `pescheria` o `pescatore` ed altri).

- **Estensione per variazioni**

L'estensione per variazioni, o somigilanza di termini, comporta la selezione di termini alternativi a quelli impostati secondo questo criterio: il termine alternativo ci considera rispondente se presenta non più di `n` errori su `m` caratteri. Automaticamente, se la dimensione del termin sul quale si sta facendo il test è un multiplo di `m`, saranno accettati un corrispondente multiplo di `n` errori.

Il concetto di errore, in questo caso, si intende come: un carattere in più, un carattere in meno o un carattere diverso.

Ecco che termini come `pastone` e `pastore` sono considerati somiglianti accettando sino ad un errore ogni 7 caratteri mentre termini come `pastura` e `pa store` richiedono 2 errori su 7 caratteri (o 1 su 4).

Anche in questo caso entra in gioco la limitazione sulle Wild Card nel prefisso della chiave e la condizione d'errore verrà applicata a partire dal primo carattere che segue la dimensione del prefisso imposta.

Si presta per risolvere principalmente problematiche inerenti errori di battitura ma anche per termini in lingua straniera se pure in modo meno accurato.

- **Estensione Fonetica**

L'estensione fonetica si presta particolarmente per risolvere ricerche su termini in lingua straniera o terminologie somiglianti (ad esempio marchi e brevetti) ed in parte anche per errori di battitura se pure solo per sostituzione o aggiunta di caratteri, non per omissione.

Prevede una configurazione per mezzo della quale esplicitare le accoppiate di sequenze di caratteri che devono avere somgilianza fonetica (es. 'PH' e 'F', oppure 'X' e 'CS'). Questa configurazione può essere personalizzata per cogliere le peculiarità delle lingue con le quali ci si trova a doversi confrontare o con le terminologie utilizzate (es. Se il rapporto tra 'PH' ed 'F' è evidente, quello tra 'Z' ed 'S' lo è meno ma è altrettanto rappresentativo).

sono inoltre previsti caratteri a scomparsa come il carattere 'H' che può essere presente o omesso. I termini selezionati in alternativa a quello espresso possono aver subito molteplici trasformazioni per ciascuno dei caratteri originari con limitazioni sull'estensione delle trasformazioni (l'estensione si applica ai primi n caratteri di un termine con n configurabile.)

- **Ricerca per range o liste di termini**

Anche in questo caso parliamo più di una restrizione della ricerca più che di una vera estensione, uno strumento di supporto alla ricerca. I dettagli sono consultabili nell'appendice B.

Dettagli finali ed espressioni articolate

Considerando un'espressione del tipo:

```
<Nome Canale> <Operatore di Uguaglianza/Disuguaglianza> <Termini di ricerca combinati tra loro>
```

come una componente elementare della ricerca, una frase di ricerca più o meno complessa può essere composta da diverse componenti elementari a loro volta combinate per mezzo dei suddetti operatori logici. Poinchè gli operatori logici hanno un particolare ordine di valutazione, si può fare uso di parentesi tonde per racchiudere qualsiasi sotto insieme della fase di ricerca e vincere l'ordine di valutazione degli operatori per condurre la ricerca nella direzione voluta.

A completamento di quanto detto sino ad ora ogni componente elementare più essere negata anteponendo l'operatore NON (forma italiana) ovvero NOT (forma internazionale) alla componente stessa. Per ragioni che sarebbe complesso spiegare in questo frangente, la condizione d'utilizzo dell'operatore di negazione esso deve essere anteposto ad una parentesi aperta e provvede a negarne tutto il contenuto.

- **Trattamento dell'operatore “Diverso” (<>)**

L'opertore Diverso è stato per lungo tempo considerato al pari del concetto di non uguale che però non soddisfa pienamente il concetto di diverso. Per diverso, infatti, può essere inteso qualsiasi documento che non abbia il valore espresso, anche in assenza di qualsiasi valore, così come può essere inteso qualsiasi documento avente un valore valido ma diverso da quello indicato. La differenza nel comportamento e nel risultato sta nel fatto che il concetto di non uguale comporta la selezione di un documento sulla base della negazione del termine ricercato e quindi verranno selezionati anche documenti che per il canale in esame non possiedono alcun valore valido ed eventualmente, neppure il canale in se.

Vediamo secondo la versione del server il comportamento dello stesso di fronte alla richiesta di diverso (<>).

- **Comportamento antecedente la versione 20.2.1.47**

Prima della versione 20.2.1.47 l'operatore di disuguaglianza viene considerato come non uguale quindi l'espressione…

```
[Canale] <> valore
```

…viene considerata equivalente a…

```
NOT ([Canale] = valore)
```

…e, venendo convertita alla suddetta forma, porta allo stesso risultato. La scelta è caduta su questo comportamento per un duplice motivo. La ricerca per negazione, avvalendosi di una logica differente, è più veloce della ricerca per disuguaglianza e non era stata approfondita la differenza tra i concetti di diverso e di non uguale.

- **Comportamento dalla versione 20.2.1.47 in avanti**

con la versione 20.2.1.47 si è compiuta una netta differenza tra i concetti di non uguale e diverso. Il primo continua ad essere espresso per mezzo della stessa forma evidenziata nel paragrafo precedente, ovvero negando l'uguaglianza del valore indicato. Questa forma comporta la selezione di qualiasi coumento che sia privo di un qualsiasi valore nel canale indicato ovvero che contenga un valore diverso da quello espresso.

Per ottenere con un'espressione di ricerca singola solo questi ultimi documenti, ovvero quelli che hanno un valore valido per il campo indicato ma differente dal valore espresso è stato introdotti un nuovo tipo di Range, un Range per negazione. Assume la forma dei normali ranges ed impone che siano presenti entrambe gli estremi, a costo di prevederli identici, e differentemente dagli altri ranges prende i valori validi esterni al range indicato anziché i valori interni. Il separatore di range in questo caso è il `punto esclamativo` (!).

Ne consegue che la forma…

```
[Canale] <> valore
```

…viene ora tradotta nella seguente forma…

```
[Canale] ={valore!valore}
```

…portando ad un risultato che può rappresentare un sottoinsieme di quello che si sarebbe ottenuto con la forma…

```
NOT ([Canale] = valore)
```

…che seleziona anche documenti privi di qualsiasi valore in quel canale.

Con questo distinguo, e consentendo comunque la ricerca in forme esplicite, il server offre un comportamento in selezione decisamente più preciso.

- **Case and Alias.**

Ogni operatore logico può essere espresso nella frase di ricerca indipendentemente dal case. Altrettanto può essere detto per gli Alias che rappresentano i nomi dei canali di ricerca mentre questi ultimi, se espressi direttamente nella forma estesa, sono case sensitive per la parte che segue la virgola.

Negli esempi dei paragrafi successivi sarà tutto più chiaro.

Sintassi di Ricerca: Canali mono e multi valore.

Quanto detto nei paragrafi precedenti si presta ad ogni tipo di ricerca. Il presente paragrafo, per contro, ha lo scopo di chiarire come esprimere al meglio le ricerche sui diversi canali e quali risultati si possono ottenere.

In primo luogo bisogna compiere un distinguo tra i canali, distinguo legato alla natura del canale stesso ed alla tipologia di chiave che esso deve esprimere.

I canali sono quindi distinti per tipologia in…

- Alfanumerici: Canali il cui contenuto più essere rappresetato da ogni tipo di termine, alfabetico o numerico, ed il trattamento del suo contenuto viene fatto sulla base delle Stringhe che lo compongono.
- Numerici: Canali il cui contenuto deve rappresentare esclusivamente cifre. Il trattamento dei contenuti è attualmente ancora di tipo alfabetico ma verrà presto convertito ad un trattamento più accurato.
- Data: Canali il cui contenuto deve rappresentare una data intesa come anno, mese e giorno. Il contenuto della componente XML che genera questo canale di ricerca può essere espresso in diversi modi ma il sever compierà le necessarie operazioni per normalizzare tale data nel formato YYYYMMDD. La stessa normalizzazione viene applicata anche in ricerca.
- File: Canali il cui contenuto deve rappresentare un nome di file eventualmente corredato da un percorso assoluto o relativo. Il server normalizza il formato del nome espresso unificando i separatori di directory indipendentemente dalla piattaforma corrente.

- **Nota:** Salvo esplicita dichiarazione nel file di configurazione d'archivio, tutti i canali vengono creati Alfanumerici.

I canali sono inoltre distinti per tipo di indicizzazione quindi sulla base delle chiavi che vengono prodotte da essi…

- Mono Valore: L'intero contenuto della componente XML (elemento o attributo) sulla base del quale viene prodotto il vocabolario di questo canale di ricerca porta alla creazione di una ed una chiave soltanto, comprensiva di ogni carattere, spazi compresi, presenti nella componente. La presenza di spazi attigui viene ridotta ad un solo spazio e gli spazi in testa ed in coda vengono rimossi.
- Multi Valore: Il contenuto della componente XML viene frazionato in singoli termini sulla base di un elenco di separatori. Ogni singolo termine isolato in questo modo porta alla costituzione di una chiave a se stante.
- A doppia indicizzazione: combina le due funzionalità producendo chiavi che si prestano per i due scopi precedenti. La componente viene quindi sottoposta ad una indicizzazione di tipo Multi Valore ed inoltre alla creazione di una chiave per il suo intero contenuto. Dal momento che queste chiavi devono convivere nello stesso vocabolario e sarebbe rischioso confonderle, le chiavi singole vengono create preponendo ad esse uno

spazio. In conclusione, quindi le chiavi prodotte per questo tipo di canali sono identiche a quelle dei canali Multi Valore e differiscono da quelle dei canali Mono Valore per lo spazio che le precede.

- **Nota:** Salvo esplicita dichiarazione nel file di configurazione dell'archivio, tutti gli elementi vengono indicizzati come Multi Valore e tutti gli attributi come Mono Valore.

Su alcuni canali, in pratica sui canali a doppia indicizzazione e Mono Valore, può essere applicata una ulteriore forma di indicizzazione detta MD5 che può condurre alla generazione di chiavi…

- MD5 Case sensitive: L'intero contenuto della componente XML viene sottoposto all'algoritmo MD5 per produrre una chiave univoca.
- MD5 Case Insensitive: L'intero contenuto della componente XML viene sottoposto all'algoritmo MD5 per produrre una chiave univoca dopo una normalizzazione a caratteri minuscoli.

- **Nota:** Questo tipo di chiave trova applicazione solitamente nella creazione di chiavi univoche per componenti XML di dimensioni consistenti, univocità che non sarebbe altresì possibile visti i limiti di dimensione delle singole chiavi. Non trova altri effettivi campi d'applicazione.

Modalità di ricerca per campi Mono e Multi Valore: uso dei doppi apici

Veniamo ora a chiarire come sia più opportuno esprimere i termini di ricerca con o senza l'uso di doppi apici.

Una chiave Multi Valore è, per definizione, ottenuta separando le singole parti di un testo sulla base di un set di separatori ne consegue che in essa tali separatori non possono essere presenti.

Nella stragrande maggioranza dei casi, quindi, una chiave ottenuta dall'indicizzazione Multi Valore di una componente XML è rappresentata da una parola secca per la rappresentazione della quale non sono richiesti doppi apici. Ad essere precisi è opportuno non farne alcun uso per motivi che verranno espressi meglio nel prossimo paragrafo.

Contrariamente a quanto detto, la produzione delle chiavi Mono Valore comprende tutti i caratteri normalmente considerati separatori che si trovassero originariamente nel componente sottoposto ad indicizzazione. La sola operazione che si compie è la normalizzazione degli spazi, vale a dire l'eliminazione di tutti gli spazi in testa ed in coda e la riduzione ad un solo spazio di tutte le sequenza di due o più spazi rilevate all'interno della chiave. Per definzione quindi, queste chiavi possono contenere dei caratteri che in fase di ricerca possono causare ambiguità o impedire di comprendere dove inizi e finisca la chiave quindi è buona norma che i termini ricercati su un canale Mono Valore vengano racchiusi tra doppi apici.

Modalità di ricerca "As Is": uso dei doppi apici nel canali Multi Valore.

A completamento del distinguo tra ricerca nei canali Multi Valore ed in quelli Mono Valore va indicata anche la ricerca As Is ovvero la ricerca esatta di una frase (tipicamente) entro un canale Multi Valore. La ricerca per adiacenza può essere espressa tra singoli canali ma assume un particolare senso (ed è più frequentemente utilizzata) per stabilire la prossimità di termini entro un singolo canale (Multi Valore ovviamente). A queto tipo d'adiacenza si aggiunge la ricerca As Is che consente di determinare la presenza di una frase esatta entro tale canale. In generale, comunque, l'uso di doppi apici in un canale testuale sottoposto ad indicizzazione Multi Valore scatena un particolare tipo di ricerca che prevede di validare il documento se e solo se il documento contiene esattamente quella sequenza di termini senza che essi siano sottoposti ad alcuna estensione ed esattemente nell'ordine indicato, senza alcun altro termine tra loro.

Appendice A: I modificatori della ricerca, generale e dei singoli canali

Esistono due tipi di modificatori della frase di ricerca. Alcuni hanno valenza sull'intera frase di ricerca ed altri solo sul singolo canale in cui vengono espressi. I primi si rappresentano come nomi di canali, anche se assolutamente finti e vengono rimossi prima dell'effettiva analisi della sintassi di ricerca, i secondo fanno parte di questa sintassi e vengono associati al canale corrispondente a mano a mano essi vengono incontrati.

Modificatori Globali della ricerca

I modificatori globali della ricerca si rappresentano con una parentesi quadra seguita da un punto interrogativo e dal nome del modificatore. Esso può essere a sua volta seguito da parametri specifici (solitamente separati dal carattere ':') e comcluso con una parentesi quadra chiusa.

- **Nota:** I modificatori generici possono trovarsi in qualsiasi punto della frase di ricerca: in testa, in coda o nel mezzo.

I modificatori globali della ricerca sono i seguenti:

- REVSORT Richiede che l'ordinamento della selezione richiesta venga ordinato al contrario rispetto alle regole indicate o semplicemente rispetto all'ordine naturale che viene imposto ai documenti. L'ordine naturale è quello in cui i documenti sono stati inseriti nell'archivio.
- FORCEOR Richiede che, indipendentemente dalla sintassi della frase di ricerca essa venga svolta come se ogni operatore logico fosse un OR. L'applicazione di questo modificatore, solitamente, è il tentativo di ripetere una selezione che non ha dato esito per trovare documenti vicini quanto più possibile a quello che si voleva ottenere una volta che la ricerca originaria è risultata eccessivamente stringente.
- ANYALIAS Nelle ricerche che vengono effettuate parallelamente su più archivi richiede che un Alias di canale che su un dato archivio non esiste non provochi errori sintattici ma consenta l'esecuzione della ricerca con un comportamento neutrale
- GLOBALDA Altresì nota come Global Document And. Richiede che nelle ricerche per Campo Chiocciola l'operatore AND si consideri nell'ambito dell'intero documento e non nell'ambito del singolo campo.
- SIMILAR Richiede che la selezione che si sta per eseguire venga naturalmente e direttametne estesa (per mezzo del canale di ricerca testo) ai socumenti simili. Gli uni e gli altri verranno a far parte del risultato finale a meno che non si indichi il sotto parametro SKIP_ORIGIN ad indicare che solo i documenti simili devono appartenere al risultato. Indicando il parametro BASEFACTOR si indica il fattore di selettività.
- PROBAB Modificatori della ricerca probabilistica ovvero di qualsiasi forma di ricerca che preveda di tornare i documenti in un particolare ranking dato dalla pesatura dei documenti stessi secondo criteri profilabili. Ogni sotto-parametro è separato dal modificatore da un ':' ed è terminato sempre da un ':'. A seguire il valore. I modificatori sono:
    - MINRANK per indicare il valore della percentuale di somiglianza al di sotto della quale non vale la pena tornare i documenti.
    - HITWEIGHT indica invece la percentuale di peso complessivo da assegnare ai successi riscontrati e non al peso (selettività) dei termini riscontrati sul documento.
    - MINRANKSIZE indica il numero minimo di documenti da tornare. Questa impostazione sovrascrive tanto il valore di default quanto l'eventuale valore configurato in nomearchivio.conf.xml.

    - RANKSIZE indica il numero massimo di documenti da tornare. Anche questo sovrascive tanto il default (100) quanto il valore indicato nel file di configurazione dell'archivio.
- PROBABREFINE: Da utilizzarsi quando la selezione comprende, tra le altre parti l'esito di una precedente ricerca probabilistica. In tal caso l'esito della selezione viene tornato nello stesso ordine e con gli stessi voti attribuiti ai documenti della ricerca probabilistica. Se la selezione non corrisponde esattamente ad un raffinamento e quindi seleziona documenti che non si trovano nella ricerca probabilistica originaria, essi vengono posti in coda alla selezione prodotta con voto '0'. Se la selezione prevede nella sua sintassi che vengano espresse più selezioni, solo quella probabilistica viene presa in esame per l'ordinamento. Se più di una è probabilistica si acquisisce la prima espressa da sinistra verso destra.
- PHONO: Richiede che la ricerca venga compiuta in estensione fonetica anche se il corrispondente bit del comando non è attivo.
- MAXDOC: Definisce il numero massimo di documenti da tornare, tale valore è l'unico parametro previsto.
- MFSP: Richiede l'attivazione dell'estensione per genere e numero (maschile, femminile, singolare e plurale) dei termini applicati a tutti i canali di ricerca.
- SINON: Richiede l'attivazione dell'estensione per termini sinonimi avvalendosi del supporto di un Thesaurus appositamente configurato.
- VAR: Richiede l'attivazione dell'estensione per termini simili dove tali termini sono identificabili come simili in quanto contenenti caratteri in più, in meno o semplicemente diversi. Questa differenza si considera un errore. La sintassi prevede tre valori, divisi dal carattere ':', atti ad indentificare:
    - Il prefisso, ovvero quanti caratteri a partire dall'inizio del termine debbano comunque risultare invariati.
    - Il numero di errori
    - Il numero di caratteri per errore
- Questo vuol dire che la sintassi [?VAR:1:1:5] indica che il prefisso dev'essere di un singolo carattere (tutti i termini devono iniziare per quello stesso carattere) e che si tollera un errore ogni 5 caratteri. In pratica, termini di 1 solo carattere non vengono sottoposti a test ed i termini con meno di 5+1 (char + prefisso) caratteri non vengono valutati. A partire dai termini di 6 caratteri è ammesso un errore (un carattere in più, uno in meno o uno differente, tranne il primo che è nel prefisso). Si accettano due errori su 11 caratteri, 3 su 15 e così via.

- **Nota:** I modificatori sono Case Insensitive. Questi modificatori si sommano comunque ai parametri di ricerca che influenzano estensioni, distanza ed ordinamento delle adiacenze, modalità di valutazione dei termini simili o dei termini con Wild Cards e così via.

Questi parametri non vengono documenti in questa sede.

Modificatori dei singoli Canali di ricerca

I modificatori del singolo canale di ricerca trovano posto entro le parentesi quadre che isolano il nome del canale stesso e sono separati da esso per mezzo del carattere '|' (pipe). A destra di tale carattere, isolati da parentesi tonde, trovano posto i modificatori il cui formato è quindi…

```
(<Nome Modificatore>[:<eventuali parametri>])
```

Un modificatore privo di parametri non presenta quindi i due punti in se. Più modificatori vengono semplicemente espressi in sequenza senza bisogno di ulteriori separatori. Ogni carattere al di fuori delle parentesi tonde verrà ignorato.

I modificatori sono:

- AdjIgnore: Indica che per il canale in esame si desidera che la ricerca venga effettuata in AND nonostante l'espressione della stessa sottintenda un'adiacenza in quanto i termini sono scritti in sequenza senza operatori booleani tra loro. Confligge con i successivi modificatori AdjDist e AdjBay.
    - **A partire da:** 19.4.1.*
- AdjDist: Indica la distanza entro la quale accettare le adiacenze tra i termini ricercati sul canale. Come unico parametro prevede il numero che rappresenta tale distanza, eventualmente negativo se non si richiede di rispettare anche l'ordine dei termini ma solo la posizione.
- AdjBay: Indica come unico parametro il canale che rappresenta il bacino di pescaggio.
- SrcStp: Indica, come unico parametro, il nome del file di stoplist da utilizzare in ricerca. L'uso di una stoplist non cambia sensibilmente il risultato della ricerca. Ha impatto nelle adiacenze e nell'uso delle varie particelle della lingua italiana (congiunzioni, preposizioni semplici ed articolate, ecc. ecc.) che potrebbero essere assenti nei testi. Indicando null si inibisce l'uso della stoplist altrimenti utilizzata per default.
- Dll_SearchTranslate: Indica l'uso della libreria dinamica d'archivio, appositamente creata, perché valuti l'insieme dei termini cercati sul canale e li sottoponga ad una pre lavorazione che può modifiare la struttura finale della ricerca effettuata.

- **Nota:** Il Case dei modifiatori è Sensitive. Esistono modificatori tesi all'uso dell'estensione linguistica (linguaggio naturale, indicizzazione e ricerca per concetti e non per parole) propria dell'estensione delle funzionalità del server per mezzo del supporto del software prodotto da Expert System. Questi modificatori richiedono una configurazione considerevole dell'archivio e non vengono quindi trattati in questa sede.

Appendice B: Esempi

Per poter fare efficaciemente alcuni esempi immaginiamo di avere un archivio che abbia, tra gli altri, i seguenti canali:

- CognomeENome: Alias del canale XML,/documento/@cogn_nome corrispondente ad un attributo che viene sottoposto a doppia indicizzazione.
- Numero: Alias del canale XML,/documento/@num indicizzato secondo il default. Essendo un attributo (prefisso @) esso è Mono Valore.
- Comune: Alias del canale XML,/documento/@comune indicizzato secondo il default. Essendo un attributo (prefisso @) esso è Mono Valore.
- Mansione: Alias del canale XML,/documento/mansione corrispondente ad un elemento indicizzato, come da default, come Multi Valore
- Note: Alias del canale XML,/documento/note corrispondente ad un elemento indicizzato anch'esso come Multi Valore.
- @: Come sottinsieme dell'archivo rappresentato dalla Mansione e dalle Note.

Vediamo ora alcuni esempi.

La ricerca…

```
[CognomeENome]=Mario
```

…è destinata ad eleggere tutti i documenti che fanno riferimento a persone avente il nome 'Mario' in quanto il canale è stato indicizzato con una doppia indicizzazione e quindi ne sono stati estratti anche i singoli termini. Se la ricerca è stata fatta con estensione per Genere e Numero (quindi maschile /femminile/singolare/plurale) si selezionarà anche 'Roberta'.

La stessa ricerca, sullo stesso campo, ma eseguita con con i doppi apici attorno al nome indicato…

```
[CognomeENome]="Mario"
```

…troverà solo i documenti con 'Mario' e mai quelli con 'Roberta' indipendentemente dal tipo di estensione poiché la ricerca viene effettuata con la modalità As Is.

Se poi la ricerca viene effettuata indicando il blank come primo carattere della chiave cercata…

```
[CognomeENome]=" Mario"
```

…la ricerca non darà, presumibilmente, esito a meno che non ci siano documenti nei quali è stato indicato solo il nome e non il cognome. La ricerca viene infatti effettuata con la chiave completa futto della doppia indicizzazione, chiave che, come detto in precedenza, rappresenta l'intero contenuto del campo.

Maggior successo avremo ricercando…

```
[CognomeENome]=" Rossi Mario"
```

… che trova il documento con tale nome purché sia scritto in quest'ordine.

Se il campo viene alimentato anteponendo il nome al cognome la precedente ricerca non da esito mentre la ricerca…

```
[CognomeENome]=Rossi Mario
```

…che cerca in adiacenza i termini troverà il record se l'adiacenza non viene richiesta in ordine e non ne troverà in altri casi.

Per essere certi che la ricerca trovi il nostro documento con questi termini indipendentemente dall'ordine dell'adiacenza si può esprimerla come…

```
[CognomeENome|(AdjDist:-2)]= Rossi Mario
```

… oppure …

```
[CognomeEnome]= Rossi AND Mario
```

…che giungono al risultato in due modi diversi. Nel caso di adiacenza si richiede una distanza massima di due termini senza rispetto dell'ordine (in quanto il valore è negativo) ma teoricamente, nel nostro caso, anche il valore '-1' sarebbe stato sufficiente.

Se una ricerca simile viene effettuata su un canale Mono Valore, per contro, la presenza dei doppi apici assume un significato diverso. Ricercando…

```
[Comune]=Bologna
```

…si esprime una ricerca pienamente regolare in quanto il termine ricercato non presenta alcun separatore che potrebbe renderlo ambiguo mentre la ricerca…

```
([Comune]=San Lazzaro di Savena)
```

…nonostante le parentesi (che assumono valore solo per gli operatori logici) è sintatticamente errata (o quanto meno priva di significato) in quanto, una volta assunto il termine 'San' si ha subito un altro termine (condizione che sottintende l'adiacenza). Richiedendo di porre in adiacenza, quindi quanto meno in AND, due chiavi su un campo monovalore (per il quale si calcola una sola chiave per documento) si esprime una richiesta inperfetta sin dall'origine (a meno che l'attributo 'Comune' non possa essere presente più volte entro la stessa unità Informativa.

Molto più corretta e sensata sarebbe una ricerca così conformata…

```
([Comune]="San Lazzaro di Savena") OR ([Comune]="S.Lazzaro di Savena")
```

…la cui forma può essere espressa parimenti anche come…

```
([Comune]="San Lazzaro di Savena" OR "S.Lazzaro di Savena")
```

… essendo i termini relativi lo stesso canale.

Esprimendo la ricerca…

```
[Note]=Computer OR [Notte]=Calcolatore
```

… si produce un errore in quanto il canale 'Notte' non è noto. Per rendere il server tollerante verso una simile sintassi (ad esempio qualora il canale 'Notte' fosse proprio di un altro archivio) si deve esplicitare il modificatore che interviene sugli alias…

```
[?ANYALIAS] [Note]=Computer OR [Notte]=Calcolatore
```

…che da a questo punto un esito corretto.

Nelle ricerche possono poi trovare posto anche le Wild Cards. Tornando agli esempi precedenti, assumendo che il canale venga alimentato sempre nello stesso modo e volendo tutte le persone che si chiamano Rossi come cognome e che hanno il nome che inizia con 'R' possiamo esprimere la ricerca…

```
[CognomeENome]=" Rossi R*"
```

…che, nonostante l'apparenza, è molto diversa da…

```
[CognomeENome]= Rossi AND R*
```

…ovvero da…

```
[CognomeENome]= Rossi AND [CognomeENome]=R*
```

…in quanto questi due ultimi esempi troverebbero anche cognomi che inizano con la 'R' se il canale può essere presente più volte nel documento.

Per sopperire alle condizioni di ordine di compilazione si può anche esprimere la ricerca come…

```
[CognomeENome]= " Rossi R*" OR [CognomeENome]=" R* Rossi"
```

…condizione che soddisfa entrambe i casi.

Tornando però alla ricerca sulla chiavi Multi Valore la ricerca…

```
[CognomeENome|(AdjDist:-1)]= Rossi R*
```

…sfruttando l'adiacenza in disordine dovrebbe dare lo stesso identico risultato.

Copiendo una ricerca probabilistica come nell'esempio che segue…

```
[CognomeENome] = Rossi AND [Comune]=Bari AND [Note]=legge della fisica quantistica animali da cortile
```

…ci si aspetta di trovare documenti nei quali sia presente almeno uno dei termini indicati in uno dei canali indicati senza che gli operatori AND assumano il significato solitamente attribuito loro. Ciò avviene in funzione del fatto che si ipotizza che una ricerca precedentemente eseguita che non abbia dati gli esiti attesi possa essere eseguita nuovamente in forma probabilistica.

Diversamente da quanto detto, la ricerca pesata è una ricerca pura e semplica in cui l'ordinamento dei documenti avviene in base alla rappresentatività dei termini trovati in essi. Va da se che se la ricerca non prevede almeno un OR non ha gran senso eseguire una simile ricerca. Se la precedente selezione fosse espressa come…

```
[CognomeENome] = Rossi OR [Comune]=Bari OR [Note]=legge della fisica quantistica animali da cortile
```

… la ricerca pesata avrebbe senso.

Oltre a quanto indicato esiste un'altra ricerca che prevede l'applicazione di una pesatura dei documenti trovati per restituirli in un ordine di utile fruizione: si tratta della ricerca per documenti simili. Data una frase di ricerca, l'estensione ai documenti simili trova tutti i documenti che somigliano ai contenuti (canale testo) dei documenti trovati sommandoli alla ricerca appena eseguita o sostituendosi al risultato (ovvero tornando solo i documenti simili e non gli originali). Questi due comportamenti si possono ottenere con ricerche come le seguenti…

```
[Note]=testo di legge [?SIMILAR]
```

```
[Note]=testo di legge [?SIMILAR:SKIP_ORIGIN]
```

Per indicare un canale differente dal canale testo è disponibile, dalla versione 17.1.0.*, un modificatore KEYNAME da indicare come segue…

```
[Oggetto]=testo di legge [?SIMILAR:KEYNAME:Oggetto]
```

Inoltre si può indicare secondo quale criterio di selettività si debbano eleggere i termini da utilizzare per la ricerca dei documenti simili. Per definizione si eleggono termini che siano selettivi in misura maggiore del rapporto 1/100 dove 100 rappresenta l'indice di selettività. Ciò significa che la chiave, per essere di nostro ineresse, deve selezionare meno di un centesimo dei documenti in archivio. Se si vuole modificare quest'impostazione si può indicare un diverso divisore. Se ad esempio indichiamo…

```
[Oggetto]=testo di legge [?SIMILAR:BASEFACTOR:25]
```

…vogliamo indicare che una chiave è rappresentativa quando seleziona meno di un venticinquesimo dell'intero archivio. In questo modo abbiamo allargato le maglie ed avremo un maggior numero di chiavi utilizzate per la selezione. Ciò, comunque, avviene automaticamente per passi (e quindi livelli di selettività) successivi se il rapporto 1/100 non determina chiavi.

In tutte le ricerche che prevedono la pesatura dei documenti, ovver il ranking, si possono esprimere parametri che limitano o modificano il comportamento dell'algloritmo di ordinamento.

Nell'esempio che segue…

```
[Note]=testo di legge [?SIMILAR:SKIP_ORIGIN] [?PROBAB:MINRANK:40] [?PROBAB:HITWEIGHT:45]
```

…si richiede, oltre alla ricerca dei documenti simi a quelli che presentano le parole testo e @ legge in adiacenza nel canale Note, di limitare il risultato ai documenti il cui grado di somiglianza rispetti agli originali non sia inferiore al 40%. Si richiede inoltre di valutare il peso dei documenti per il 45% sulla base del numero di termini oggetto della selezione per documenti simili che è stato effettivamente rilevato nei documenti simili e per il restante 55% sulla base della selettività di tali termini. La selettività dei termini si intende tanto maggiore quanto piccolo è il numero totale di documenti che tale termine seleziona nell'archivio. Si potrebbe quindi dire che tanto più un termine è raro, tanto più esso è selettivo. Nell'esempio che segue…

```
[Note]=testo di legge [?SIMILAR:SKIP_ORIGIN] [?PROBAB:MINRANKSIZE:5] [?PROBAB:RANKSIZE:50]
```

…si vuole invece garantirsi di avere non meno di 5 documenti ma non più di 50 nel set di documenti rilevati dalla ricerca per documenti simili.

Sintassi di Ricerca Globale e Campo Chiocciola

Parlando di ricerca globale avremo alcuni casi molto diversi.

La ricerca…

```
Mario
```

…non indica alcun canale, neanche il campo chiocciola, quindi viene estesa a tutti i canali disponibili (CognomeENome, Comune, Mansione e Note) trovando la parola in uno qualsisi di questi punti. La ricerca…

```
[@]=Mario
```

…si applica invece solo ai campi Mansione e Note che sono il sottoinsieme rappresentato dal campo chiocciola. A questo punto la ricerca…

```
[@]=Mario and Rossi
```

…comporta la selezione dei documenti in cui la condizione di AND espressa sia valida in uno dei singoli campi del sottoinsieme (o la Mansione o le Note) ma non se il documento presenta sì i due termini ma uno in un canale e l'altro in un canale diverso. Per ottenere invece il documento anche in questo caso la ricerca può essere espressa come…

```
[@]=Mario and Rossi [?GLOBALDA]
```

…che imponendo l'AND di Documento consente la selezione.

Sintassi di Ricerca per Range ed elenchi di valori

Particolare attenzione va prestata alla ricerca per Range e per elenchi di valori. La ricerca per Range, infatti, può sopperire e sostituire molte (se non tutte le ricerche per disuguaglianza spuntando, solitamente, risultati migliori.

La ricerca per Range si esprime impostando il valore da ricercare come combinazione di due estremi compresi o meno. La rappresentazione si apre con una parentesi graffa aperta immediatamente seguita dal valore minore (se previsto dalla ricerca) ed eventualmente racchiuso tra doppi apici se necessario per la natura del campo e della ricerca in corso di esecuzione. Questo primo valore viene seguito dal separatore '|' (pipe) qualora gli estremi di ricerca siano compresi ovvero dal separatore '^' (accento circonflesso) qualora gli estremi di ricerca siano esclusi. Di seguito si deve porre il secondo valore e concludere il tutto con una parentesi graffa chiusa. La ricerca prevede che uno dei due estremi possa mancare ma non che possano mancare entrambe.

E' inolrte stato concepito un separatore di range che ne inverte il significato. Tale separatore è il punto esclamativo (!) che vuole intendere una negazione ovvero che la ricerca va compiuta per tutti i termini esterni al range indicato. Per definizione, in questo caso, gli estremi sono compresi nell'elenco da escludere e quindi esclusi dalla ricerca.

La sintassi impone che entrambe gli estremi siano presenti anche a costo di averli idenici se si intende ricercare tutti i valori tranne quell indicato.

In sostanza, quindi, la forma…

```
[campo]={val!val}
```

…corrisponde in tutto e per tutto alla forma…

```
[campo]<>val
```

- **A partire da:** Il separatore di inversione è statè inserito dalla verisone 20.2.1.*

Veniamo a qualche esempio chiarificatore.

La ricerca…

```
[Numero]={10|100}
```

…prevede la selezione di tutti documenti in cui il canale Numero assume un valore compreso tra 10 e 100. Tale ricerca corrisponde formalmente a…

```
[Numero]>=10 AND [Numero]<=100
```

Sull'onda di questo esempio avremo…

```
[Numero]={|10}
```

…per cercare documenti con un qualsiasi numero minore o uguale a 10, identico quindi all'espressione…

```
[Numero]<=10
```

Identico ragionamento va fatto per la ricerca per Maggiore o Uguale.

Se, per altro, la ricerca viene espressa come…

```
[Numero]={20^}
```

…la ricerca richiede ogni valore superiore a 20, quindi 20 escluso, mentre non avendo indicato l'estremo superiore non si esclude alcun valore in tale direzione. La ricerca è quindi equivalente a…

```
[Numero]>20
```

Poichè gli estremi vengono esclusi se si fa uso del separatore '^', un campo contenente solo numero interi non darebbe alcun esito per la ricerca…

```
[Numero]={14^15}
```

…in quanto gli estremio verrebbero entrambe esclusi.

Allo stesso modo la ricerca…

```
[Numero]={10|10}
```

…è un modo assolutamente arzigogolato per dire…

```
[Numero]=10
```

- **Nota:** Risulta evidente che la ricerca per Range, prevedendo un estremo inferiore ed uno superiore, non porta ad alcun esito (o provoca errore sintattico) qualora l'estremo inferiore indicato risulti essere maggiore dell'estremo superiore o qualora, come detto, mancassero entrambe gli estremi.

In estensione di quanto detto, come precedentemente annunciato, esiste il range negativo quindi la ricerca…

```
[Anno]={1990!1999}
```

…troverà i documenti nei quali appare un anno che non rientra tra il 1990 ed il 1999 entrambe compresi.

Si noti quindi che la forma…

```
[Anno]={1990!1990}
```

…che indica che non si vogliono i documenti dove appare un valore diverso da 1990 corrisponde alla forma…

```
[Anno]<>1990
```

…e rappresenta il modo in cui il server, dalla versione indicata, traduce questo tipo di espressione.

- **Attenzione:** Si noti che l'espressione “[Anno]<>1990” che ora viene tradotta in “[Anno]={1990!1990}” veniva precedentemente trasformata nella forma “(NOT([Anno]=1990))” che però non da un comportamento corretto qualora il canale sul quale si compie la ricerca è ripetibile nel documento.

C'è infatti un fondamentale distinguo tra cercare i documenti in cui `non appare` un valore, caso del NOT, e cercare i documenti in cui `appare` un valore `diverso da`, condizione che non preclude, se il canale di ricerca è ripetuto nel documento, che il valore che si intende negare sia effettivamente presente in un'altra ripetizione.

Alla ricerca per Range di valori si affianca quelle per un elenco dato. Sintatticamente molto simile, l'elenco dei valori, separati da virgole ed eventualmente racchiusi tra doppi apici, deve essere sempre espresso tra parentesi graffe finendo col rappesentare una ricerca in OR in modo particolare.

La ricerca…

```
[CognomeENome]=" Rossi Mario" OR " Verdi Luca"
```

…potrebbe quindi essere espressa anche come…

```
[CognomeENome]={" Rossi Mario"," Verdi Luca"}
```

… conducendo allo stesso risultato.

Altrettanto dicasi per una ricerca del tipo…

```
[Numero]={1,3,5,7,8,2,4}
```

…nella quale si evidenza che l'ordine dei componenti l'elenco non è singificativo per un buon esito della ricerca.