SISTEMA DI CLASSIFICAZIONE

NICOLA MORETTO (MATR. 578258)

21 settembre 2012

Il documento presenta i risultati delle fasi di analisi e di progettazione dei nuovi criteri di classificazione.

Versione	Data	Modifiche
0.1	10-09-2012	Prima stesura del documento.
0.2	11-09-2012	Aggiunto il capitolo contenuti informativi.
0.3	12-09-2012	Aggiunto il capitolo requisiti.
0.4	13-09-2012	Ampliato il capitolo requisiti.
0.5	14-09-2012	Rivisto il capitolo requisiti.
1.0	15-09-2012	Pubblicazione della prima versione ufficiale.
1.1	18-09-2012	Rivista e ampliata la sezione requisiti.
1.2	19-09-2012	Aggiornate le sezioni *Entità* e *Etichette*.
1.3	21-09-2012	Aggiornata la sezione *Contenuti*.

Tabella 1: Registro delle modifiche

INDICE

1 CONTENUTI INFORMATIVI

1.1 INTRODUZIONE

Il patrimonio di conoscenza della piattaforma è garantito essenzialmente dai contenuti pubblicati dagli utenti, che condividono alcune proprietà essenziali (autore, data di pubblicazione, visibilità, ...) e un contenuto informativo vero e proprio, di lunghezza (massima) variabile.

Classi

Le classi di contenuti rispecchiano altrettante forme di espressione quotidiana (la domanda, il pensiero elementare, un discorso articolato, ...), facilmente riconoscibili da qualsiasi utente, e di contenuto (audio, video, evento, ...).

1.2 CRITERI DI CLASSIFICAZIONE

Per facilitare la catalogazione e il reperimento dei contenuti, essi condividono, a prescindere dalla rispettiva classe, i medesimi criteri di classificazione, ciascuno dei quali ne valuta e pesa un aspetto differente:

1.2.1 *Argomento*

Branca del sapere - agnostica rispetto al tema specifico della piattaforma - entro la quale ciascun contenuto della piattaforma si colloca univocamente.

1.2.2 *Emozione*

Emozioni personali che l'autore associa al contenuto al momento sdella redazione.

1.2.3 *Intenzioni*

Intenzioni con cui l'autore scrive il contenuto (opinione, critica, ...) e utili a chiarire lo spirito con cui debba essere interpretato.

1.2.4 *Giudizi*

Giudizi qualitativi espressi dagli altri utenti su un contenuto. I criteri e i parametri con cui tali valutazioni verranno espresse sono attualmente in fase di indagine da parte di altri membri del team di progetto.

1.3 CLASSI

1.3.1 *Documento*

La classe DOCUMENTO è concepita per esprimere un contenuto prevalentemente testuale, di lunghezza rilevante e articolato nella struttura; al suo interno l'utente può esporre delle tesi o opinioni, supportandole con opportune argomentazioni, notizie dettagliate, ...

1.3.2 *Domanda*

La classe DOMANDA offre la possibilità di sottoporre agli utenti della piattaforma una domanda relativa ad un certo tema o ad un contenuto specifico.

1.3.3 *Evento*

La classe EVENTO permette di pubblicizzare un evento o manifestazione, indicandone luogo e data di svolgimento, se sia pubblico o privato, ...

1.3.4 *Multimedia*

La classe MULTIMEDIA consente di pubblicare contenuti audio e video, sia in risposta sia in forma completamente autonoma rispetto ad altri contenuti informativi.

1.3.5 *Pensiero*

La classe PENSIERO è concepita per esprimere idee, concetti o pensieri semplici ed essenziali, la cui lunghezza risulta dunque limitata.

1.3.6 *Risposta*

La classe RISPOSTA offre la possibilità di inserire una risposta ad una domanda precedente o un commento ad un generico contenuto.

1.4 RELAZIONI

Contenuto generico

All'interno della piattaforma il generico contenuto riveste un ruolo essenziale rappresentando l'astrazione fondamentale su cui poggiano tutti i tipi di contenuti e sulla quale è definita la maggior parte delle relazioni, sia interne (tra i contenuti stessi) sia esterne (criteri di classificazione, ...).

Discussione

A ciascun contenuto pubblicato nella piattaforma è possibile rispondere con altri del medesimo tipo o differente: ciò implica che, a partire da un contenuto qualsiasi, può nascere una discussione in grado di svilupparsi e ramificarsi con il massimo grado di libertà, non essendovi limiti sui tipi di contenuti o sul tema.

Ad esempio, una risposta ad un contenuto può - in virtù di una particolare associazione di idee - riguardare un tema non strettamente correlato al contenuto di partenza.

2

REQUISITI

Conoscenza

Ove la conoscenza della piattaforma è generata dai contenuti pubblicati dagli utenti, si rende necessario un criterio (o insieme di criteri) di classificazione per facilitare e rendere più efficienti possibili la *catalogazione*, il *reperimento* e la *consultazione* delle informazioni in essi contenute.

Classificatore

Il classificatore tiene traccia dei frammenti di informazione presenti nei contenuti, ciascuno dei quali può riferire una o più ENTITÀ del dominio della piattaforma; nella sua essenza, il criterio di classificazione deve quindi provvedere ad associare a ciascun contenuto delle ETICHETTE, che contrassegnano le entità citate al suo interno.

2.1 ENTITÀ

Entità

Le ENTITÀ della piattaforma rappresentano elementi concreti (luoghi, persone, ...) o astratti (concetti, ...) a cui afferiscono i contenuti.

Dizionario

Il DOMINIO della piattaforma rappresenta l'insieme di entità definite - in un dato instante - all'interno della piattaforma e risulta, per certi versi, paragonabile ad un dizionario linguistico, costituito da una insieme di lemmi, ciascuno dei quali possiede svariati significati (ACCEZIONI), identificanti - a seconda del contesto - altrettante entità del dominio.

2.1.1 *Identificazione univoca*

Sinonimi

Gli utenti possono in genere riferire la stessa entità (concreta o astratta) mediante termini o espressioni differenti: tale ambiguità linguistica rappresenta un ostacolo imprescindibile ma cruciale per un'identificazione chiara e consistente di ciascuna entità all'interno della piattaforma e rende di conseguenza più complesso stabilire se due o più contenuti riferiscano la medesima entità.

Ambiguità sintattica

Ciascuna entità del dominio della piattaforma richiede perciò di essere identificata in modo univoco da un termine o un'espressione al fine di eliminare possibili ambiguità sintattiche e renderla così riferibile e riconoscibile - dall'utente o dal sistema - in modo consistente all'interno di qualsiasi contenuto.

Sintassi e semantica

In caso contrario, una conseguenza immediata sarebbe una minore accuratezza dei risultati di ricerca, dovuta alla restituzione dei soli contenuti nei quali l'entità sia identificata precisamente dall'etichetta scelta. L'esito desiderato consisterebbe invece nell'insieme di contenuti in cui l'entità in questione sia riferita, a prescindere dalla specifica etichetta utilizzata: in altre parole, si desidera che la ricerca venga trasferita dal piano puramente sintattico (l'etichetta specifica) a quello semantico (l'entità indicata dall'etichetta).

$$\text{entità} \rightarrow \text{identificatore}$$

2.1.2 *Identificazione non ambigua*

Accezioni

Ciascun termine o espressione può assumere significati differenti (ACCEZIONI) - e dunque identificare entità distinte - a seconda del contesto in cui è inserito o citato.

Riveste un'importanza cruciale poter stabilire senza ambiguità all'interno di ciascun contenuto a quale accezione del termine o dell'espressione si faccia riferimento, per consentire una corretta identificazione dell'entità riferita. *Ambiguità semantica*

$$\text{identificatore} \rightarrow \text{entità}$$

2.1.3 *Gestione delle relazioni*

Osservando la similitudine tra il dominio delle entità e un dizionario linguistico, si nota immediatamente l'esistenza di relazioni gerarchiche (dal generale al particolare) tra le entità, che si traducono nella possibilità di associare a ciascuna entità un numero arbitrario di padri (entità generiche) e figli (entità specialistiche).

Ciascuna entità ha $0 \ldots n$ *figli*

Ciascuna entità ammette naturalmente delle sotto-entità specialistiche, che ne rappresentano un aspetto o sfaccettatura particolare.

Ciascuna entità ha $0 \ldots n$ *padri*

A differenze della struttura gerarchica classica, ove ciascun elemento può avere molti figli ma un solo padre, il dominio delle entità estende la relazione *uno-a-molti* anche agli elementi padre per consentire di esprimere l'eventuale ambiguità associata ad una generica entità, ossia la possibilità che essa trovi collocazione logica in diverse posizioni all'interno della gerarchia. *Padri e figli*

Principio di sostituzione

Il principio di sostituzione implica l'esistenza di relazioni nascoste, frutto dell'ereditarietà gerarchica e particolarmente rilevanti nella selezione di contenuti riguardanti una determinata entità: essa va infatti estesa ricorsivamente a tutte le entità figlie di quella data.

2.1.4 *Ricerca di un'entità*

La ricerca di un'entità da parte dell'utente risulta facilitata dalla struttura gerarchica, che consente attraverso un processo dicotomico (dal generale al particolare) di portarla a termine nel modo più efficiente possibile. Per ulteriori informazioni, consultare la sezione 2.3.2.

2.2 ETICHETTE

Riprendendo il modello concettuale accennato nella sezione 2.1, può risultare conveniente immaginare il dizionario D come l'unione di sottoinsiemi E_i, ciascuno dei quali corrisponde ad un'entità distinta e contiene esattamente un'ETICHETTA PRIMARIA e_0, che identifica univocamente il sottoinsieme/entità in questione, e gli eventuali SINONIMI e_j (in numero arbitrario, anche nullo).[1][2] *Entità ed etichette*

1 $i \in \mathbb{N}, i \leq n = |D|$
2 $j \in \mathbb{N}, j \leq m = |E_i|$

2.2.1 *Gestione dei sinonimi*

Copertura sintattica

Sebbene ciascuna entità sia identificata univocamente da un'etichetta primaria all'interno di qualsiasi contenuto, i sinonimi vengono memorizzati e conservati nel dizionario poiché rivestono un ruolo altrettanto cruciale: dal momento che ciascun utente può cercare o riferirsi ad un'entità non solo mediante il suo identificatore univoco (l'etichetta primaria) ma anche tramite una qualsiasi forma alternativa ma semanticamente equivalente (un sinonimo), conservare questi ultimi consente di individuare con maggior probabilità e precisione l'entità cui l'utente fa riferimento, di stabilire se essa sia già definita all'interno del dominio della piattaforma e di aggiungere eventualmente il termine o l'espressione cercata come nuova etichetta (primaria o sinonimica).

Ciascuna etichetta può avere $0 \dots n$ *sinonimi*

Come accennato in precedenza, è possibile riferirsi ad un'entità con termini o espressioni differenti, sebbene all'interno della piattaforma l'identificazione sia univoca e dunque tutti i sinonimi rimandino ad una precisa e specifica etichetta primaria.

Etichette primarie e sinonimiche

Per evitare la proliferazione di etichette duplicate (sintatticamente differenti ma riferenti la medesima entità), che contribuirebbe a indebolire l'efficacia (qualità dei risultati di ricerca, navigabilità dei contenuti, ...) e l'efficienza (dimensione del dizionario, ...) del sistema di classificazione, risulta utile, per ogni entità E_i:

1. definire un'etichetta che la identifichi chiaramente all'interno della piattaforma (ETICHETTA PRIMARIA e_0);
2. tenere traccia dei sinonimi utilizzati dagli utenti per riferire tale entità (ETICHETTE SINONIMICHE e_j).

Aggiunta di un sinonimo ad un'etichetta

Ogni qualvolta un utente suggerisce una nuova etichetta e, che risulti sinonimo di un'altra esistente $e_j \in E_i$, essa viene aggiunta al dizionario interno della piattaforma come $e_{m+1} \in E_i$ sinonimo di $e_0 \in E_i$; da quel momento, qualora un utente provi ad assegnarla ad un contenuto della piattaforma, il sistema assegnerà automaticamente la corrispondente etichetta primaria e_0.

Accezioni e sinonimi

Non si da il caso che la nuova etichetta e_{m+1} possa essere sinonimo - rispetto ad una specifica accezione - di due (o più) etichette primarie, ma può essere sinonimo di etichette primarie in numero al più pari alle relative accezioni.

Uno-a-molti

Si considerino ad esempio due etichette primarie, $e_1 \in E_i$ e $e_2 \in E_i$: per la proprietà transitiva, se e_1 è sinonimo di e_{m+1} e e_2 è sinonimo di e_{m+1}, allora e_1 e e_2 sono a loro volta sinonimi; ma allora, in accordo ai principi sopra illustrati, l'ultima tra e_1 e e_2 ad essere stata aggiunta doveva essere inserita nel sottoinsieme dell'altra, contraddicendo così le ipotesi iniziali.

Eliminazione di un sinonimo associato ad un'etichetta

In considerazione delle esigenze di copertura sintattica, l'eliminazione di un sinonimo associato ad un'etichetta avviene solo in condizioni molto particolari, tali da invalidare la relazione sinonimica tra l'etichetta primarie e il sinonimo stesso.

2.2.2 *Gestione delle accezioni*

Ciascuna etichetta può avere $1 \dots n$ *accezioni*

Ambiguità semantica

Ciascuna etichetta può riferirsi a entità differenti a seconda del contesto, perciò diventa indispensabile poterne precisare le possibili accezioni $a_k \in A$.[3]

Accezioni, entità e sottoinsiemi

Con l'introduzione delle accezioni, il dizionario della piattaforma acquisisce una nuova dimensione poiché ciascuna etichetta - al variare dell'accezione - si riferisce ad un'entità differente e può essere:

PRIMARIA
: L'etichetta identifica univocamente un'entità del dominio e ha un numero arbitrario di sinonimi.

SINONIMICA
: L'etichetta rappresenta un sinonimo di un'etichetta primaria.

Aggiunta di un'accezione ad un'etichetta

L'aggiunta di un'accezione ad un'etichetta consiste nel definire il contesto o ambito in cui essa assuma un significato univoco e non equivocabile.

Eliminazione di un'accezione associata ad un'etichetta

L'eliminazione di un'accezione $a_k \in A_j$ associata ad un'etichetta $e_j \in E_i$ prevede due possibili casi:

ETICHETTA PRIMARIA:
: Se l'etichetta è primaria, l'accezione viene eliminata e un sinonimo viene promosso in sua vece ad etichetta primaria.

ETICHETTA SINONIMICA
: Se l'etichetta è sinonimica, si procede direttamente alla cancellazione dell'accezione.

2.2.3 *Gestione del dizionario*

Il dizionario contiene in ogni istante

$$\sum_{i \in \mathbb{N}, i \leq n} |E_i|$$

etichette, a ciascuna delle quali sono associate $\left|A_{i,j}\right|$ accezioni.

Il dizionario contiene $0 \dots n$ *etichette*

Il dizionario contiene un numero di etichette almeno pari al numero di entità definite poiché ciascuna entità dev'essere identificata dalla corrispondente etichetta primaria:

$$\sum_{i \leq n} |E_i| \geq \sum_{i \leq n} min\{|E_i|\} = \sum_{i \leq n} 1 = n$$

Inserimento di una nuova etichetta

L'aggiunta di un'etichetta primaria implica l'identificazione di una nuova entità non ancora presente nel dizionario, l'assegnazione dell'etichetta primaria come identificatore univoco e l'inserimento nella gerarchia.

3 $k \in \mathbb{N}, k \leq t = |A|$

Eliminazione di un'etichetta esistente

L'eliminazione di un'etichetta $e_j \in E_i$ richiede di considerare separatamente ogni possibile accezione $a_k \in A_j$, valutando caso per caso:

ETICHETTA PRIMARIA
: Se l'etichetta è primaria viene eliminata e un sinonimo viene promosso in sua vece ad etichetta primaria.

ETICHETTA SINONIMICA
: Se l'etichetta è sinonimica si procede direttamente alla cancellazione.

2.3 CONTENUTI

2.3.1 *Gestione delle etichette*

Catalogazione dell'informazione

Le etichette primarie rappresentano lo strumento essenziale per identificare e tracciare le entità riferite all'interno di un contenuto.

A ciascun contenuto possono essere assegnate $0 \dots n$ *etichette*

Ciascun contenuto può citare o fare riferimento a svariate entità al suo interno, perciò dev'essere possibile assegnargli diverse etichette primarie, in numero pari e corrispondenti alle entità in questione.

Assegnazione di un'etichetta ad un contenuto

L'assegnazione di un'etichetta ad un contenuto consiste nell'individuazione di parole o brevi espressioni chiave, che identifichino un'entità concreta (luogo, persona, oggetto, ...) o astratta (concetto, argomento, ...) riferita o citata all'interno del contenuto stesso.

Etichetta esistente

Una volta individuata, il sistema deve verificare se essa sia già stata utilizzata in precedenza (e quindi già presente nel dizionario interno). In caso affermativo, può trattarsi di:

ETICHETTA PRIMARIA
: L'etichetta viene associata al contenuto.

ETICHETTA SINONIMICA
: Al contenuto viene assegnata la corrispondente etichetta primaria.

Nuova etichetta

In caso contrario, viene indagata la presenza nel dizionario interno di etichette sintatticamente equivalenti a quella immessa dall'utente. La ricerca può presentare due possibili esiti:

NESSUN RISULTATO
: La parola o espressione viene memorizzata nel dizionario come etichetta primaria.

ETICHETTA PRIMARIA
: La parola o espressione viene memorizzata nel dizionario come sinonimo dell'etichetta primaria.

Al termine della procedura viene assegnata in entrambi i casi al contenuto un'etichetta primaria, rispetto alla quale l'utente è chiamato a specificare - ove disponibili in numero maggiore di uno - un'accezione.

Eliminazione di un'etichetta associata ad un contenuto

La rimozione di un'etichetta assegnata in precedenza ad un contenuto non altera in alcun modo il dizionario interno, anche qualora essa non risultasse assegnata ad altri contenuti.

2.3.2 *Ricerca e navigazione*

Reperimento dell'informazione

La ricerca e la consultazione dei contenuti rappresentano attività cruciali per gli utenti della piattaforma e ci si affida al criteri di classificazione delle etichette per reperire in maniera efficiente le informazioni cercate.

L'approccio e lo scopo con cui gli utenti navigano l'insieme di contenuti disponibili all'interno della piattaforma può tuttavia differire sensibilmente.

Ricerca di contenuti generici

Gerarchia

L'utente interessato a conoscere gli argomenti discussi nella piattaforma procede in genere ad esplorare i contenuti partendo dalle entità, per facilitare la cui navigazione si definisce una struttura gerarchica (dal generale al particolare), che le raccoglie e le cataloga in maniera ordinata (v. sezione 2.1.3).

Dicotomia

Tale soluzione permette all'utente di esplorare in maniera più efficiente il dominio delle entità e individuare i contenuti di interesse, afferenti ad una specifica entità/etichetta primaria, più rapidamente grazie ad un PROCESSO DICOTOMICO.

RICERCA DI UN'ETICHETTA L'utente alla ricerca di informazioni su un particolare tema inizia con l'individuare le etichette aventi maggiore attinenza e rilevanza. La ricerca di corrispondenze nel dizionario prevede che:

1. vengano prese in esame tutte le etichette $e \in E_i$, poiché solo contemplando le chiavi primarie e i relativi sinonimi si massimizza la probabilità di ottenere riscontri positivi (maggiore copertura sintattica);
2. vengano restituite le chiavi primarie corrispondenti alla ricerca;
3. per ogni sinonimo $e_j \in E_i$ individuato, si restituisce la corrispondente etichetta primaria $e_0 \in E_i$.

Ricerca di contenuti specifici

Etichette e accezioni

La ricerca di informazioni su un tema specifico viene effettuata specificando una o più etichette, eventualmente declinate nelle specifiche accezioni, che presentino agli occhi dell'utente particolare attinenza e siano dunque con maggior probabilità associate ai contenuti di interesse.

Insiemi di etichette

Siano E_s l'insieme delle etichette cercate e E_c l'insieme delle etichette assegnate ad un generico contenuto: il primo passo consiste nel sostituire le etichette sinonimiche con le equivalenti primarie ed estendere l'insieme E_s alle etichette figlie di ogni $e \in E_s$.

Corrispondenza

A questo punto si possono distinguere tre casi principali, a seconda del grado di corrispondenza/attinenza dei contenuti rispetto alle etichette cercate:

CORRISPONDENZA COMPLETA: $E_s \subseteq E_c$
: Al contenuto risultano assegnate tutte le etichette richieste dall'utente (massima attinenza).

CORRISPONDENZA PARZIALE: $E_s \cap E_c \neq \emptyset$
: Al contenuto risulta assegnata parte delle etichette richieste dall'utente (media attinenza).

NESSUNA CORRISPONDENZA: $E_s \cap E_c = \emptyset$
: Al contenuto non risulta assegnata alcuna etichetta richiesta dall'utente (attinenza nulla).

Attinenza

I contenuti attinenti possono essere visualizzati in ordine decrescente rispetto al numero di etichette assegnate corrispondenti a quelle richieste dall'utente:

$$|E_s \cap E_c|$$

Ricerca di contenuti affini

La ricerca di contenuti affini consiste nell'identificare, a partire da un contenuto dato, altri la cui pertinenza sia massima: in questo scenario valgono le medesime considerazioni emerse nella sezione precedente, previa sostituzione di U_e con l'insieme delle etichette assegnate al contenuto corrente.