Social Tagging as a Classification and Search Strategy

Advisors:
FARINETTI Laura
KHOKHAR Ashfaq

Candidate:
BONINO Serena

Settembre 2007
Summary

In recent years, tagging systems have become increasingly popular. These systems enable users to add keywords (tags) to Internet resources (e.g. web pages, images, videos) without relying on a controlled vocabulary. Tagging systems have the potential to improve search and personal organization while introducing new modalities of social communication and opportunities for data mining. This potential is largely due to the social structure that underlies many of the current systems.

This thesis analyzes these systems. Specifically, the work is organized in two main parts: the first one involves a statistical analysis of tags to infer semantics of tag aggregations and the second one is experimental and devoted to discover some semantic aggregation between resources and tags.
Contents

Summary

1 Riassunto 1

1.1 Introduzione 1

1.1.1 Sistemi di social tagging 2

1.2 Cause ed Obiettivi 3

1.2.1 Definizione del problema 4

1.2.2 Soluzione proposta 5

1.3 Sviluppo 6

1.3.1 Dati sperimentali 6

1.3.2 Caso di studio 9

1.3.3 Fase sperimentale 23

1.4 Conclusioni 25

2 Introduction 27

3 Rationale 30

3.1 Causes 30

3.2 Problem Statement 31

3.3 Proposed Solution 32
4 State of Art

4.1 Why do People Tag? ............................................. 35
  4.1.1 From solitary to social .................................. 35
  4.1.2 Ad-hoc groups or crowds ................................. 36
  4.1.3 Four conditions for wisdom of crowds ................. 36
  4.1.4 Tagging and Collaborative Filtering .................... 37

4.2 Social Tagging and Social Bookmarking Systems .......... 37

4.3 Tagging and Taxonomy ........................................ 40

4.4 Tagging Systems Problems ................................... 43
  4.4.1 Polysemy ..................................................... 43
  4.4.2 Synonymy ..................................................... 44
  4.4.3 Basic Level .................................................. 44

4.5 Social Tagging Tools .......................................... 47
  4.5.1 Closed source platforms ................................ 50
  4.5.2 Open source platforms ................................... 52
  4.5.3 Final Consideration ........................................ 60

5 Case Study: Connotea Data Analysis and Search Improvements 28

5.1 Connotea ....................................................... 28

5.2 The Data ....................................................... 31
  5.2.1 Users ......................................................... 32
  5.2.2 Resources .................................................... 33
  5.2.3 Tags ......................................................... 33
  5.2.4 Tagged objects ............................................. 33

5.3 Statistical Analysis ........................................... 35
  5.3.1 Users activity and Tag Quantity ......................... 35
  5.3.2 Resources .................................................... 35
  5.3.3 Kinds of Tags ............................................... 35
5.4 Experimental Work ................................. 37
  5.4.1 FreeTAG ........................................ 37
  5.4.2 Tag Stemming ................................... 41
  5.4.3 WordNet ......................................... 50

5 Results Analysis ................................. 61
  5.1 Statistical Analysis Results ......................... 61
    5.1.1 Users activity and Tag Quantity .................. 61
    5.1.2 Resources ......................................... 67
    5.1.3 Kinds of Tag .................................... 71
  5.2 Experimental Work Results .......................... 73
    5.2.1 Tags ............................................. 73
    5.2.2 Resources ......................................... 81

6 Conclusions ......................................... 103

Bibliografía ............................................ 105
Chapter 1

Riassunto

1.1 Introduzione

Uno dei punti fondamentali dell’evoluzione del Web verso il Semantic Web è rappresentato dai sistemi di ricerca semantici. Tra le diverse funzionalità fornite da questi sistemi si possono evidenziare la ricerca, l’aggregazione e la combinazione dell’informazione. Questi obiettivi possono essere raggiunti considerando tre elementi fondamentali: la diversità dei media, la varietà delle sorgenti e la non omogeneità delle strategie di ricerca. Un sistema che riesca ad operare soddisfacciendo tutti questi requisiti attribuisce alle applicazioni un livello di versatilità e flessibilità che non è possibile raggiungere basandosi esclusivamente sulle informazioni (knowledge base).

L’obiettivo fondamentale di questo lavoro di tesi è fornire una soluzione per l’aggregazione semantica delle parole chiave e delle risorse coinvolte in un sistema di social tagging al fine di migliorare il meccanismo di ricerca delle risorse in questo contesto. I risultati presentati sono stati ottenuti considerando diversi casi di studio.

La prima parte di questo riassunto è volto a presentare il contesto in cui si è operato, i sistemi di social tagging; segue una presentazione dettagliata degli obiettivi del lavoro e delle soluzioni proposte; infine sono presentati i risultati ottenuti...
attraverso la verifica sperimentale.

1.1.1 Sistemi di social tagging

Nata nel 1950 per scopi militari, la rete Internet ha raggiunto il pubblico a partire dagli anni '90 fornendo un ingente numero di servizi quali posta elettronica, motori di ricerca, file sharing, streaming e possibilità di collaborazione tra utenti diversi.

Il costo trascurabile e la possibilità di scambiare idee, conoscenze e capacità in modo quasi istantaneo, ha reso sempre più facile il lavoro di équipe, in cui le persone si riuniscono e collaborano attraverso una comunicazione basata sull’utilizzo di computer formando delle comunità on-line.

Un tipo implicito di questo tipo di lavoro è rappresentato dai sistemi di social tagging. In questi sistemi gli utenti caratterizzano i contenuti attraverso tag. Un tag è una parola chiave, un termine, che gli utenti associano ad un pezzo di informazione (rappresentata da una risorsa Internet, da una fotografia, da un articolo o ancora da un video clip), e che permette la descrizione di un concetto e la possibilità di una classificazione personalizzata basata su parole chiave. Attraverso l’operazione di tag delle risorse, gli utenti caratterizzano quest’ultime memorizzando i loro riferimenti come bookmark personali. La collezione di tag, risorse e tagger (le persone che taggano) sono condivisi. In questo modo persone con interessi simili possono trovare collegamenti a risorse per categoria attraverso un’operazione di ricerca basata su tag.

L’aumento della popolarità e della competizione tra i sistemi di social tagging ha condotto all’introduzione di nuovi servizi che esulano dalla pura e semplice condivisione dei bookmark. Questi riguardano, tra gli altri, la possibilità di inserire commenti, l’abilità nell’importare, esportare ed inserire appunti e la creazione di gruppi e reti sociali. Molti dei suddetti sistemi hanno anche implementato alcuni algoritmi che suppongono una correlazione tra parole chiave, utenti o risorse, attraverso l’esame delle relazioni tra i tag e del clustering di questi ultimi.
A causa della rapida espansione di applicazioni che supportano le operazioni di tagging delle risorse, questo tipo di sistemi non è ancora stato studiato in modo approfondito e capito nei dettagli. Proprio questo fatto rende sempre più necessarie analisi e studi volti all’introduzione di miglioramenti ai sistemi già esistenti.

Partendo da queste considerazioni, questo lavoro, basato sullo studio di Connotea, un sito web che supporta il tagging di risorse web relative a pubblicazioni accademiche, ha lo scopo di valutare la possibilità di aggregazione tra tag e tra risorse.

1.2 Cause ed Obiettivi

La crescita costante, in questi ultimi anni, del numero di utenti che utilizzano i sistemi di social tagging, ha causato un’evoluzione di questi ultimi che sono passati dal fornire un semplice modo per l’organizzazione delle informazioni al prevedere una suddivisione delle conoscenze tra persone diverse. Nonostante questa necessità di collaborazione e di cooperazione, ci sono ancora molti individui che lavorano in un modo totalmente autonomo ed individuale. Il fatto che gli utenti taggino le risorse in modo disattento, solitario ed asociale sottolinea l’emergere di alcuni problemi quali la presenza di errori di spelling, di tag con più di un significato, di parole chiave non chiare a causa della confusione tra sinonimi e contrari, etc. Ulteriori problemi derivano dal fatto che alcuni utenti utilizzano metodi di tagging personalizzati e non sempre ‘ortodossi’ e che non esiste ancora un meccanismo che possa essere utilizzato per indicare le relazioni gerarchiche tra diverse parole chiave. Questo ultimo aspetto può essere spiegato meglio per mezzo del seguente esempio: una risorsa potrebbe venire etichettata sia come formaggio oppure fontina, senza che ci sia un modo per indicare che la fontina è un particolare tipo (una sotto classe) di formaggio. Nonostante tutto, l’elevato numero di persone coinvolte in questi tipi di sistemi contribuisce a mitigare gli effetti di questi errori e distrazioni.
1.2.1 Definizione del problema

Gli utenti che classificano i contenuti e generano i metadati per mezzo di sistemi online per la condivisione delle conoscenze, hanno evidenziato due problemi principali [10]:

- la struttura delle taxonomie o dei metadati potrebbero essere così rigide da non supportare le necessità degli utenti: in questi sistemi gli utenti devono taggare le risorse con un ben definito insieme di parole chiave che richiede loro una quantità di lavoro aggiuntiva siccome si rende necessaria l’operazione di verifica della consistenza con la taxonomia data per ogni parola chiave inserita. Questo limita la libertà di caratterizzazione delle risorse;

- il costo aggiuntivo rappresentato dalla classificazione pesa sugli utenti, ma la collettività gode dei benefici: spesso l’operazione di tagging è ancora vissuta come un’esperienza individuale. Gli utenti preferirebbero sentirsi completamente liberi di usare qualunque parola chiave essi trovino rappresentativa, senza tenere in considerazione i tag precedentemente inseriti da altre persone.

Il tentativo di risolvere questi problemi ha spinto i ricercatori a studiare e migliorare i sistemi che supportano il social tagging.

Il primo problema converge ad una ‘classificazione sociale’ generata dagli utenti: coloro che non non sono in grado di inserire nulla più che una semplice parola chiave potrebbero descrive e collegare i loro metadati. Lo svantaggio immediato è la perdita di precisione: non viene eseguito alcun controllo sui sinonimi e sui contrari; i termini correlati non sono considerati e il contesto generale non viene preso in considerazione.

Il secondo problema evidenzia la necessità di identificare la classificazione che emerge dalle operazioni di tag collettivo, specialmente quando il numero di utenti, tag e risorse diventa grande. Infatti, spesso accade che solo un esiguo numero di
tag per ogni risorsa diventi ‘popolare’ rispetto agli altri sinonimi considerati nelle fasi iniziali. Sfortunatamente, quando il sistema è popolato da un numero ridotto di risorse, utenti e tag, diventa più difficile identificare un insieme di tag popolari siccome questi ultimi rimangono confusi tra gli altri sinonimi. Un tentativo volto a ridurre questo fenomeno e a velocizzare la convergenza si basa sulla ricerca di metodi adatti all’aggregazione di parole chiave che potrebbero essere in qualche modo correlate semanticamente e che, pertanto, potrebbero essere rappresentate per mezzo di un singolo tag.

In questo lavoro sono proposti alcuni casi di studio che sono sviluppati per verificare la possibilità di raggiungere questi obiettivi.

1.2.2 Soluzione proposta

Questo lavoro di tesi include un’analisi statistica di un insieme di dati volta alla caratterizzazione delle parole chiave. Lo studio statistico è suddiviso in più parti che differiscono per gli aspetti considerati: la prima parte esamina il comportamento degli utenti, la seconda analizza i diversi tipi di tag, mentre la terza consiste in una caratterizzazione delle risorse.

Tutte queste considerazioni sono utili per chiarire quale sia la distribuzione dei tag, degli utenti e delle risorse all’interno dell’insieme di dati utilizzato.

Sostanzialmente viene proposto un approccio per l’aggregazione delle parole chiave basato sul loro significato semantico con lo scopo di migliorare il ritrovamento delle risorse nei sistemi di social tagging. Secondo questo approccio la confusione tra sinonimi e contrari dovrebbe essere ridotta e potrebbero essere forniti alcuni standard per la struttura delle parole chiave. In particolare, alcuni degli errori più comuni verrebbero eliminati: per ridurre le imprecisioni legate alle forme plurali e singolari delle parole viene utilizzato un algoritmo di stemming che riduce la parola
alla sua radice; mentre, per eliminare i problemi di lettere maiuscole minuscole si utilizza una funzione di normalizzazione che riporta tutti i caratteri in minuscolo, senza spazi né caratteri speciali. Inoltre, il problema della dispersione dovuta ai sinonimi e ai contrari viene evitata per mezzo dell’utilizzo di un dizionario semantico per la lingua inglese che è in grado di individuare l’insieme di sinonimi di una parola data.

Successivamente viene proposto un altro tipo di aggregazione: l’idea, essenzialmente, è quella di sfruttare i tag inseriti liberamente dagli utenti per collegare le risorse a cui essi sono associati. In altre parole, quando più persone utilizzano lo stesso (sotto)insieme di tag, le risorse associate ad essi sono considerate correlate a livello semantico. Quindi, una ricerca per parola chiave restituisce come risultato tutte le risorse che sono state in parte o completamente etichettate con l’insieme di tag considerato: le risorse aggregate in questo modo potrebbero facilitare in modo molto consistente le operazioni di ricerca, ritrovamento e navigazione nel tempo.

1.3 Sviluppo

Dopo aver presentato i requisiti e gli obiettivi che questa tesi intende raggiungere, si passa ora alla descrizione della fase di sviluppo. In questa sezione, viene descritta brevemente l’organizzazione dei dati che sono stati utilizzati e vengono presentati i casi di studio considerati al fine di valutare la validità del lavoro da un punto di vista qualitativo.

1.3.1 Dati sperimentali

La necessità di avere a disposizione un set di dati da analizzare ha portato alla ricerca di una piattaforma che supporti il social tagging che abbia le seguenti caratteristiche:

- un numero sufficiente di utenti
un account personale

- un meccanismo per inserire collegamenti, titoli e descrizioni

- una documentazione sufficientemente estesa per guidare l’utente, comprensiva di FAQ e pagine di help

- la possibilità di inserire anche più di una parola chiave alla volta

- una classificazione tra metodi di tagging ‘libero’ o ‘aperto’.

I risultati ottenuti e le caratteristiche delle piattaforme trovate sono presentati nelle tabelle 4.1 e 4.2. La differenza principale tra i risultati ottenuti è che alcuni di questi tool sono open source, mentre altri no.

Proprio questa differenza ha condotto alla scelta della piattaforma da utilizzare in questo lavoro di tesi. Infatti, tra le varie alternative, si è scelto un tool open source con le suddette caratteristiche che fornisce API utili nella fase di download dei dati: Connotea.

I dati raccolti sono relativi al periodo compreso tra il 9 dicembre 2004 ed il 28 novembre 2006 e consistono in un insieme di utenti, uno di risorse ed un’altra di parole chiave. Essi sono stati memorizzati in un database organizzato nel seguente modo (fig. 1.1):

La tabella degli utenti (user) è stata riempita con i dati ottenuti per mezzo della API www.connotea.org/data/users e contiene 23852 tuple.

Per la tabella delle risorse (resource) è stata utilizzata la API www.connotea.org/data/bookmarks, che ha restituito 242943 risultati.

I dati scaricati, relativi ai tag (API utilizzata www.connotea.org/data/tags) sono stati sufficienti a riempire solo alcuni dei campi della tabella finale (tabella tag). La cardinalità della tabella è di 81392 tuple.
Figura 1.1. Organizzazione del database

<table>
<thead>
<tr>
<th>USER</th>
<th>RESOURCE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>user_id</strong></td>
<td><strong>username</strong></td>
</tr>
<tr>
<td>1</td>
<td>MMQSballoon</td>
</tr>
<tr>
<td>2</td>
<td>ajurgens</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TAG</th>
<th>TAGGED_OBJECTS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>tag_id</strong></td>
<td><strong>download tag</strong></td>
</tr>
<tr>
<td>4</td>
<td>librarian+competencies</td>
</tr>
</tbody>
</table>
Le informazioni contenute nelle tre precedenti tabelle sono state unite per mezzo della API www.connotea.org/data/date/YYYYMMDD in una nuova tabella (tagged_object) composta da 275274 righe.

1.3.2 Caso di studio

Una volta raccolti i dati, il lavoro è stato suddiviso in due parti: la prima sezione relativa ad un’analisi statistica volta a capire se la loro distribuzione fosse significativa e pertanto potessero essere utilizzati per ottenere i risultati sperimentali significativi, e la seconda relativa all’analisi delle possibilità di aggregazione tra tag e tra risorse.

Analisi statistica

L’analisi statistica dei dati si focalizza su quattro punti fondamentali:

- attività degli utenti
- numero di tag
- risorse
- tipi di tag

Attività degli utenti

In questa fase si vuole analizzare il numero di parole chiave presenti nelle liste dei tag dei diversi utenti e la loro distribuzione.

Nel set di dati analizzato è interessante notare la relazione molto forte che esiste tra l’”età” dell’account dell’utente (cioè il periodo di tempo intercorso dalla creazione dell’account) ed il numero di giorni in cui gli utenti hanno inserito almeno un bookmark. Come si può notare dalla figura (5.1), nell’insieme di utenti considerato ve ne sono molti che non hanno mai inserito alcuna parola chiave. Tipicamente questi
utenti sono quelli ‘più recenti’, cioè quelli che si sono appena registrati e non hanno ancora inserito alcun tag.

Un altro aspetto interessante è rappresentato dalla figura 5.3 che evidenzia come alcuni utenti utilizzino un vasto insieme di parole chiave, mentre altri no. In altri termini, si può notare che al crescere del numero di bookmark, il numero di tag utilizzati cresce ma con tasso di crescita molto differente: alcuni utenti usano molte parole diverse per caratterizzare le risorse, altri tendono ad usare spesso le stesse parole.

**Numero di tag**

In questa fase si vuole analizzare il numero di parole chiave distinte utilizzate per etichettare le risorse.

La figura 5.11 mostra che ci sono poche risorse taggate con molti tag differenti e che in generale nel sistema sono utilizzate molte parole chiave diverse. Questo risultato fa emergere la necessità di uno studio sulla possibile aggregazione dei tag da utilizzarsi come punto di collegamento tra le risorse.

**Risorse**

In questa parte dell’analisi si vuole vedere come varia nel tempo l’inserimento dei bookmark.

Tipicamente le URL ricevono la maggior parte dei bookmark molto velocemente ed il tasso di inserimento decresce nel tempo: quindi, molto spesso, le risorse raggiungono la ‘popolarità’ molto in fretta, nei primi giorni.

Come dimostrato dalle figure 5.6, 5.7, 5.8 e 5.9, in un solo caso, l’ultimo, la risorsa ha bisogno di un periodo di tempo maggiore per diventare popolare. In più, le figure precedenti mostrano che il numero di tag associati alle risorse è ridotto. Questo comportamento è molto strano e giustificabile considerando che le risorse considerate spesso sono relative ad argomenti molto specifici e quindi un numero esiguo di persone è interessato ad esse.
**Tipi di tag**

Per caratterizzare i diversi tipi di parole chiave viene utilizzato il median rank come mostrato in figura 5.10. Esso è calcolato come funzione inversa della frequenza dei tag: i tag più frequenti sono quelli caratterizzati dai valori più piccoli di median rank.

In generale, gli utenti che utilizzano i sistemi di tag collettivo hanno comportamenti molto diversi per quanto riguarda le operazioni di tagging: alcuni hanno molti tag, altri pochi.

La prevalenza di risorse caratterizzate da un grande numero di parole chiave e le informazioni intrinseche ottenute dall'analisi dei tagger dimostra che la maggior parte delle operazioni di tagging sono fatte per uso personale. Nonostante questo, le risorse taggate per uso personale potrebbero essere utili per altri utenti.

Partendo da questo punto è possibile supporre che le risorse taggate potrebbero essere valutate sia in forma aggregata che individualmente.

**Lavoro sperimentale**

Il punto di partenza del lavoro sperimentale è rappresentato da tre domande:

- E’ possibile trovare relazioni concettuali tra risorse e tag differenti?
- Le risorse taggate con le stesse parole chiave possono essere raggruppate in qualche modo?
- Quando due o più risorse/tag possono essere considerati simili?

Lo studio sviluppato al fine di dare una risposta a questi quesiti si basa sui seguenti approcci:

- tag normalizzati
- tag stemmati
• sinonimi dei tag

**Tag Normalizzati**

Nello sviluppo di questa tecnica è stato utilizzata una libreria open source che implementa funzionalità sociali: FreeTAG. Questa piattaforma fornisce molte funzionalità tra le quali la normalizzazione dei tag, la ricerca di tag ed oggetti simili, la lista delle risorse caratterizzate da una specifica parola chiave, la lista dei tag più popolari e delle risorse più recenti, etc.

È importante definire cosa si intende per normalizzazione. Quando si parla di tag, la normalizzazione riguarda tutte le operazioni che modificano/filtrano gli input inseriti dagli utenti quali: la trasformazione di tutte le lettere in minuscolo, l’eliminazione degli spazi, l’eliminazione della punteggiatura, etc. FreeTAG normalizza tutti i tag di default, eliminando tutti i caratteri non alfanumerici e rendendo tutte le lettere minuscole.

L’utilizzo della funzione di FreeTAG `normalize_tag` porta ad un aggiornamento del database (fig. 1.2).

Successivamente viene utilizzata la funzione `similar_objects` al fine di trovare le risorse correlate. Questo metodo in linea di principio cerca, nel database, le risorse che potrebbero essere simili ad una data attraverso un matching tra gli oggetti che condividono una o più parole chiave.

I risultati forniti sono caratterizzati da:

• un peso ⇒ numero floating point che rappresenta quanto due risorse siano correlate e può assumere un valore compreso tra 0-1.0

• un `object_id` ⇒ ID univoco dell’oggetto correlato.
Figura 1.2. Organizzazione del database

<table>
<thead>
<tr>
<th>USER</th>
<th>RESOURCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>user_id</td>
<td>username</td>
</tr>
<tr>
<td>1</td>
<td>MMQSBalloon</td>
</tr>
<tr>
<td>2</td>
<td>ajurgens</td>
</tr>
</tbody>
</table>

| TAG | TAGGED OBJECTS |
|----------------|
| tag_id | download tag | raw tag | tag | stemmed tag |
| 4      | librarian+competencies | librarian competencies | librariancompetencies |
| 6      | stress+granules | stress granules | stressgranules |
| tag_id | tagger_id | tagged_on | object_id |
| 1      | 275      | 2006-11-27 15:11:25 | 1337 |
| 2      | 275      | 2006-11-27 15:11:25 | 1337 |
La rappresentazione concettuale di questa tecnica è fornita in figura 5.1 e il procedimento seguito può essere facilmente compreso leggendo il seguente estratto di pseudo-codice:

selezionare una risorsa e cercare i tag associati
{
    per ogni tag
    {
        cercare e collezionare tutti i tag normalizzati associati ad essa
        per ogni tag normalizzato
        {
            cercare e collezionare tutti quei tag che hanno questa forma normalizzata;
            per ognuno di questi tag
            {
                mostrare le risorse associate
            }
        }
    }
}
Figura 1.3.  FreeTAG approach

**Tag Stemmati**

Questa sezione è dedicata alla presentazione dell’approccio basato sull’utilizzo di un algoritmo di stemming.

Molto spesso, le parole chiave considerate hanno simili significati semantici e potrebbero pertanto essere organizzate in modo da rendere esplicita questa equivalenza. Per questo motivo, un certo numero di algoritmi di stemming, anche detti stemmer, sono stati sviluppati. Il loro obiettivo è quello di ridurre le parole alla loro radice. La conseguenza immediata consiste in una riduzione della dimensione del dizionario, cioè del numero di termini diversi necessari per rappresentare un insieme di documenti. I vantaggi derivanti dalla riduzione della dimensione del vocabolario sono rappresentati dalla minore quantità di memoria necessaria e dal minore tempo di processamento.

Esistono molte tecniche di stemming tra le quali alcuni approcci ibridi. Questi tipi di approcci usano più tecniche diverse contemporaneamente.

Nel lavoro presentato viene utilizzato uno di questi metodi di stemming, l’algoritmo di Porter. L’idea di fondo di questo algoritmo consiste nella constatazione che i suffissi della lingua inglese (approssimativamente 1200) sono costituiti dalla combinazione di suffissi più piccoli e semplici. Esso è composto da cinque passi in ognuno dei quali si applicano una serie di regole: se una regola su di un suffisso è verificata,
la condizione specificata da quella regola viene testata sul risultato dell’operazione di stem (il suffisso viene rimosso secondo le modalità previste dalla regola).

Una volta soddisfatta la condizione di una regola, il suffisso viene rimosso e si procede con il passo successivo. Se la regola non è accettata si verifica la condizione successiva fino a quando non si trova una regola che venga soddisfatta o non ci siano più regole definite per quel passo. A questo punto si passa al passo successivo. Il processo viene iterato per tutti e cinque i passi (fig. 5.2).

L’applicazione dell’algoritmo di stemming al database considerato implica una ulteriore modifica dell’architettura come presentato in figura 1.4.

![Diagram](image.png)

Figura 1.4. Organizzazione del database

Una volta ottenuti la forma stemmata dei tag si procede alla modifica delle funzionalità di FreeTAG (*freetag.class.php*) al fine di utilizzare quest’ultima forma invece di quella normalizzata.
Figura 1.5. Retrieval of similar resources through stemmed tags
La rappresentazione concettuale di questa tecnica è fornita in figura 5.3 e il procedimento seguito può essere facilmente compreso leggendo il seguente estratto di pseudo-codice:

```pseudo-codice
selezionare una risorsa e cercare i tag associati ad essa
{
   per ogni tag normalizzato
   {
      cercare e collezionare tutti i tag stemmati;
      per ogni tag stemmato
      {
         cercare e collezionare tutti i tag che hanno quest’ultima come forma stemmata;
         per ogni tag trovato
         {
            mostrare le risorse associate
         }
      }
   }
}
```

Teoricamente, il set dei tag normalizzati è un sottoinsieme di quelli stemmati così che

- • il numero di risorse ottenute in quest’ultimo caso è maggiore

- • è possibile trovare risorse che non sono correlate a causa dell’inesattezza dello stemmer
non vengono presi in considerazione i sinonimi dei tag.

**Sinonimi dei tag**

La tecnica che considera i sinonimi dei tag è stata realizzata per mezzo dell’utilizzo di WordNet.

WordNet è un dizionario semantico per la lingua inglese. Esso raggruppa le parole inglesi in insiemi di sinonimi chiamati synset, fornisce alcune definizioni brevi e generali e memorizza le relazioni semantiche tra questi set di sinonimi.

L’attitudine di WordNet nel trovare i sinonimi delle parole è stata utilizzata in questa tesi. Più nel dettaglio la funzione `getSynsetWithoutSenseDisambiguation` viene utilizzata per il ritrovamento dei synset senza disambiguazione di senso. Essa considera tutti i sinonimi senza verificare che siano effettivamente correlati tra loro. I risultati ottenuti attraverso questa operazione hanno reso necessaria un’ulteriore modifica all’architettura della base dati come mostrato in figura 1.6.
L'obiettivo, a questo punto, riguarda la possibilità di trovare risorse simili tra loro attraverso l'uso dei sinonimi. In particolare si vogliono trovare tutte le risorse che sono state taggate con almeno uno dei sinonimi associati ai tag di una data risorsa (5.4).
Figura 1.7. Retrieval of similar resources through synonymous tags
Il procedimento seguito durante la fase di ricerca viene spiegato attraverso il seguente pezzo di pseudo-codice:

```
per ogni risorsa
{
    per ognuno dei tag
    {
        cercare e collezionare tutti i sinonimi;
        per ogni sinonimo
        {
            cercare e collezionare i tag che hanno questo sinonimo nel proprio synset;
            per ogni tag collezionato
            {
                mostrare le risorse associate
            }
        }
    }
}
```

Il risultato che ci si aspetta a questo punto è di avere un maggiore numero di risorse considerate simili perché di fatto l’insieme restituito da *FreeTag* è un sottoinsieme di quest’ultimo set. Sfortunatamente, questo approccio potrebbe introdurre del ‘rumore’: infatti le risorse associate al sinonimo di un tag non sono necessariamente relative allo stesso contesto.
1.3.3 Fase sperimentale

Durante la fase di test sono state considerate due situazioni differenti:

- correlazione tra tag e risorse

- correlazione tra risorse diverse.

**Correlazione tra tag e risorse**

Durante questa fase sono stati considerati 7 diversi tag appartenenti all’insieme delle parole chiave più popolari. Su questi tag sono stati implementati i tre approcci descritti sopra e, come si aspettava, il numero di risorse restituite cresce a partire dalla tecnica basata sui tag normalizzati, fino ad arrivare a quella basata sui sinonimi.

Successivamente, per ognuna delle parole chiave considerate, si esaminano le prime 10 risorse restituite dai tre approcci e si analizzano i risultati ottenuti attraverso una valutazione manuale. Come mostrato dalle figure 5.15, 5.16 e 5.17, si possono avere quattro possibili tipi di valutazione:

- positiva nel caso in cui la parola chiave sia menzionata esplicitamente nel titolo o nel corpo della risorsa

- media quando la risorsa è chiaramente collegata al tag

- negativa in tutte le altre situazioni

- in alcuni casi le risorse ritornate non erano disponibili o perché non più presenti in rete oppure perché richiedevano un’operazione di autenticazione per l’accesso; a questi casi è stata associata un’etichetta di ‘non trovata’.

E’ importante notare che la correlazione tra tag e risorse è strettamente dipendente dalla ‘generalità’ delle parole chiave: più i tag sono specifici, migliori sono i risultati ottenuti.
Correlazione tra risorse diverse

Questa fase di test riguarda l’applicazione delle tre tecniche analizzate in precedente a 17 risorse scelte in maniera del tutto casuale.

Anche in questo caso è possibile notare l’incremento del numero di risultati restituiti da ognuno dei tre approcci considerati.

Per ognuna delle 17 risorse esaminate si analizzano i primi 10 risultati restituiti da ognuna delle precedenti tecniche e si attribuisce una valutazione in base alle seguenti regole:

- se le risorse restituite sono relative allo stesso argomento la valutazione è buona
- quando gli argomenti trattati dalle risorse restituite sono solo parzialmente correlati la valutazione è media
- nel caso in cui gli argomenti trattati siano completamente diversi la valutazione è negativa
- la valutazione ‘non trovata’ viene assegnata in tutti i casi elencati precedentemente.

I risultati ottenuti sono mostrati nelle figure 5.18, 5.19 e 5.21. Da questi grafici è possibile evincere alcuni dei problemi che interessano questo caso di studio: nell’approccio basato sui tag normalizzati è definito un meccanismo di ranking per le risorse restituite, lo stesso meccanismo è adottato nella tecnica basata sui tag stemmati, per l’approccio basato sui sinonimi non è fornito alcuni meccanismo di ranking e questa potrebbe essere la causa della scarsa qualità dei risultati ottenuti.

Per verificare che questa supposizione sul degrado delle performance, lo stesso esperimento descritto precedentemente è stato applicato a 17 risorse scelte tra quelle che avevano un insieme di risultati restituiti con cardinalità non superiore a 9.

I metodi di valutazione dei risultati sono gli stessi utilizzati precedentemente, e i risultati ottenuti (fig. 5.25, 5.26 e 5.27) mostrano che effettivamente è necessario
trovare modi più efficaci per la valutazione del ranking dei risultati nel caso dei tag stemmati e della tecnica basata sui sinonimi.

1.4 Conclusioni

L'obiettivo principale di questo lavoro di tesi consiste nell’analisi della possibilità di aggregazione tra tag e tra risorse per uno specifico sito web che supporta le funzionalità di social tagging: Connotea.

La prima parte dello studio è relativo ad un’analisi statistica dei dati collezionati per vedere se la loro distribuzione sia significativa. In particolare i risultati ottenuti sono stati confrontati con quelli ottenuti da [5] relativi al sito web Del.icio.us. Come dimostrato da questa operazione di confronto, nella maggior parte dei casi i due insiemi di dati sono confrontabili. Ne sono un esempio il numero dei tag presenti nelle tag list degli utenti (fig. 5.1) oppure il comportamento del tasso di crescita dei tag per due utenti estremi (fig. 5.3). Tuttavia esistono alcune situazioni, come il tasso di crescita di alcuni tag selezionati (fig. 5.4) oppure l’evoluzione temporale dell’inserimento dei bookmark (fig. 5.7), in cui i due insiemi di dati non sono confrontabili. Questo risultato è dovuto alla natura intrinseca dei dati: Del.icio.us fornisce un servizio web di social bookmarking volto alla memorizzazione, condivisione e scoperta di bookmark web in generale, mentre Connotea è basato su dati molto più specializzati siccome rappresenta un riferimento per ricercatori, personale clinico e scienziati. Inoltre il numero di utenti, tag e risorse disponibile su Connotea è minore di quello fornito da Del.icio.us.

Dopo l’analisi dei dati, il lavoro si è concretato sulla ricerca e sullo studio, in un primo momento di possibili relazioni tra tag e risorse taggated, e successivamente sulla possibilità di aggregazione tra risorse. Le tecniche utilizzate per questo proposito si basano sull’utilizzo di tag normalizzati, stemmati e basati su sinonimi.
La prima parte dell’analisi ha fatto emergere il fatto che le risorse restituite come risultato tipicamente sono state taggate con le parole chiave fornite, in particolare per quanto riguarda l’approccio basato sui sinonimi. I risultati ottenuti sono buoni in generale: in tutti i casi le risorse restituite (data una certa parola chiave) sono effettivamente collegate a quella fornita in input.

Questa parte dello studio è utile da un punto di vista di aggregazione concettuale dove i tag collegati a risorse simili possono essere aggregati al fine di creare un’ontologia che potrebbe portare l’utente ad un nuovo approccio per la ricerca non più basato su parole chiave, ma basato su concetti.

Inoltre, i risultati ottenuti dalla seconda parte dello studio hanno evidenziato una certa attitudine all’aggregazione tra risorse in tutti e tre i casi di studio. In particolare, si prevedeva che i risultati migliori fossero forniti dalla tecnica basata sui sinonimi. Nella realtà invece questa situazione si è verificata solo in 7 tra i 17 casi presi in considerazione. Questo risultato è accettabile se si considera il fatto che in teoria questa tecnica dovrebbe essere quella che introduce la maggiore quantità di rumore siccome considera i sinonimi di una parola data senza fare una disambiguazione di significato (quindi non viene analizzato il contesto per trovare il significato reale della parola). Tuttavia questa parte del lavoro ha sottolineato che un metodo intelligente per l’associazione di un ‘peso’ alle risorse simili debba essere introdotto in alcune situazioni (ad esempio nella tecnica basata sui sinonimi) ed in altre debba essere migliorato (ad esempio nell’approccio basato sui tag stemmati).

In conclusione questa nuova tecnica di ricerca, dopo i miglioramenti necessari, potrebbe diventare un punto di incontro fondamentale nello sviluppo e nella diffusione della nuova concezione di semantic web, sviluppatisi in questi ultimi anni.
Capitolo 2

Introduction

Born in the 1950s for military purposes, the Internet gained a public face in the 1990s and, from this moment on, its usage grew in an exponential way providing a lot of different services like e-mail, search engine, voice telephony, file sharing, streaming media and collaboration.

The low-cost and nearly instantaneous sharing of ideas, knowledge, and skills has made collaborative work, in which people rendezvous, connect or collaborate through computer-mediated communication forming on-line communities, dramatically easier.

An implicit type of collaborative work is represented by social tagging systems. In such systems, users mark content with descriptive tags. A tag is a keyword or a term which users can associate with or assign to a piece of information (e.g. an Internet resource like a picture, an article, or a video clip), describing the item and enabling customized keyword-based classifications. By tagging resources, users categorize them and store their references as personal bookmarks. The collections of tags, resources and taggers (people who tag) are shared, so that other people with similar interests can retrieve resource links by category, given the associated tags. In effect, most of these social system services allow users to search for bookmarks
associated with a set of tags, and rank the resources with the number of users who have bookmarked them.

The increasing popularity and competition of social tagging systems have extended the services to offer more than just sharing bookmarks, such as rating, commenting, the ability to import and export, add notes, reviews, email links, automatic notification, feed subscription, web annotation, creation of groups and Social Networks. Many social bookmarking services have also implemented algorithms that infer implicit correlations between tags, users or resources from the tags that are assigned to resources, by examining the clustering of particular keywords, and the relation between tags.

Despite the rapid expansion of applications that support tagging of resources, tagging systems are still not well studied or understood. This leads to study possible improvements of social tagging engines.

Our research is based on the study of Connotea, a site that allows for the collaborative tagging of shared website bookmarks related to academic publications.

We present a preliminary analysis of the previously mentioned tagging system to explore some of the issues that can be encountered in one sample system. We compare our findings to the work of Golder and Huberman [5] on Del.icio.us.

Furthermore, we use this study to outline and motivate the direction of research devoted to find out some possible aggregation between tags and resources.

One of the aspects involved in the study of social tagging platforms is to identify the emergent classification that comes out from the collaborative tagging of all the involved users, especially when the number of users, tags, and resources becomes large. In fact, the main problem consists in finding out an emergent set of popular tags that could remain blurred among synonyms, especially when the system has a small population. A possible solution to this problem can be represented by the development of some procedures intended for the aggregation of tags which seem
Another kind of aggregation can be done on resources. The idea is to analyze all the tags associated to each resource so as to put together the resources which are tagged or partially tagged with the same analyzed word. In this way when a search by tag is done, the result can involve a large amount of resources.

In this context, the main objectives of the thesis are the study and evaluation of the recent user-centered classification and search approaches like social tagging systems. Chapter 2 explains the motivations of the thesis work, in particular the causes and the real problem are pointed out and a solution is proposed. Chapter 3 depicts the current state of art of social tagging systems exploiting the available social tagging systems either commercial or open source. Follows the explanation of the motivations that led to the choice of one particular open source platform: FreeTAG. Chapter 4 describes the work organization for the collection of data, the analysis of the data set and experimentations on semantic-based aggregation of tags and resources. Chapter 5 discusses the obtained results for both the statistical and experimental works while Chapter 6 contains the conclusions and some hints for future developments.
Capitolo 3

Rationale

3.1 Causes

In the last few years a growing amount of people has used social tagging systems not only as a way to organize the information, but also as a way to share knowledge with other taggers. Notwithstanding this need of collaboration and cooperation, there are still many people that work in an independent way. As users tag in an inattentive, asocial and solitary way, the highlighting problem is the presence of errors represented by miss-tagging due to spelling errors, tags that can have more than one meaning, unclear tags due to synonym/antonym confusion, highly unorthodox and personalized tag schema from some users and no mechanism for users to indicate hierarchical relationships between tags. This last aspect can be better explained considering the following example: a resource might be labeled as both cheese and cheddar, with no mechanism that might indicate that cheddar is a refinement or sub-class of cheese. However the presence of a lot of different people involved in a social tagging system lighten such errors and distractions.

Further, even if there are such inexactitudes, platforms continue to work but not in a proper way: the process of elimination of uncorrect tags could improve
3 – Rationale

performances and this is our staff.

Next sections will better analyze the problem statement considering the actual state of art of social tagging systems and will propose some attempts of solution used in the work.

3.2 Problem Statement

Users who classify contents and generate metadata through on-line knowledge sharing systems have pointed out two main issues [10]:

- the taxonomy or metadata structure may be too rigid to support user needs: in such systems users have to tag resources with a given well defined set of keywords that asks them for more work (to check if the chosen tag is consistent with the taxonomy) and limits their freedom of resources characterization;

- the overheads of classification are borne by the user, but the group reaps the benefits: often tagging is still lived as a individual experience so users would like to feel completely free to use whatever keywords they want without taking into account the tags already used by other people.

The attempt to solve these problems have pushed researchers to study and improve social tagging engines.

The first problem converges to a user-generated social classification: users who do not have the confidence or the inclination to apply anything more than a keyword can describe and link their metadata. The obvious disadvantage is the lack of precision: synonym/antonym control is not performed, related terms are not considered and the global context is not taken into account.

The second issue spots the necessity of identifying the classification that emerges out from the collaborative tagging of all the involved users, especially when the
number of users, tags, and resources becomes large. In fact, often happens that only a few tags for each resource seem to become popular among other synonyms that some users initially exploits. Unfortunately, when fewer users, tags and resources populate the system, it is more difficult to identify an emergent set of popular tags, as they remain blurred among synonyms. An attempt to reduce this phenomenon and to quicken the convergence is to find some method to aggregate tags which are possibly semantically correlated, and therefore representable as a single tag.

Experiments in this direction are proposed in this work, starting from the next section.

3.3 Proposed Solution

In this thesis work a statistical analysis of an existing dataset is performed to analyze tags population. The statistical study is separated in different sections depending on which aspects are considered: the first section examines users behavior, the second one exploits all the different kind of tags while the third analyzes resources. In particular the analysis traces tags use of individual users, to spot differences in men activity and to understand the popularity growth rate of some selected tags, and also traces the time evolution of how much URLs are tagged: all these considerations are useful to clarify the distribution of tags, users and resources in the dataset that will be used for experiments.

Hence an approach is proposed to aggregate tags according to their semantics with the aim of improving retrieval of resources in the context of a social tagging system. In this way synonym/antonym confusion is reduced and standards for the structure of the tags are provided. Particularly, some of the most common errors are eliminated: plural/singular words disambiguation is performed using a stemming algorithm which attempts to reduce a word to its stem, or root form,
capitalization problems are also solved using a normalization function that lowercase alphanumeric characters, removing spaces and special characters. Additionally, the synonym/antonym dispersion is avoided using a semantic lexicon for the English language that retrieves the set of synonyms, yet without performing sense disambiguation, (i.e. without checking their correlations in meaning).

Furthermore, another kind of aggregation is proposed: the idea, in essence, is to exploit the user free-text tags to link the resources they are associated to. In other words, where people happen to use the same (sub)set of tags, then the associated resources are considered as semantically correlated. Therefore, a search by tag gives as result all the resources which have been completely or partially tagged with such tags: aggregated resources in this way seem to make a body of similar information increasingly easier to search, discover, and navigate over time.

Next chapter analyzes the current state of art of social tagging systems explaining why do people tag, what are social tagging and social bookmarking systems, pointing out the difference between tagging and taxonomy and, finally, highlighting what are the problems of nowadays tagging systems.
Capitolo 4

State of Art

From the beginning, the World Wide Web has been a place for group formation, to find and connect with like-minded others. This fact has helped the development of social software in which people meet and share information through computer-mediated communication and form on-line communities. Social tagging and social bookmarking systems, instead, do not represent explicit communities, but lead to ad-hoc group creation, lowering the barriers to finding like-minded others, enabling social discovery and connections.

Figura 4.1. Social interaction and conceptual transmission with tags
4 – State of Art

4.1 Why do People Tag?

For most people, tagging means sharing their own information and watching others. Even if one tags mostly to remember its own stuff, it is difficult to remain untouched by the presence of others. How tagging lets us connect with others?

4.1.1 From solitary to social

Web browsing can be a solitary experience. Computers are individualistic devices. Many afternoons, one sits at its desk, browsing the web, listening to music. He comes across a resource he wants to remember. He tags it. In that moment, he goes from wandering the web alone to joining a group of others. This transition is important. In a moment, he is transported to a crowd of people with whom he has at least one thing in common. And best of all, he can enjoy their presence, but he does not need to converse. Often, one likes to be in the company of others, without needing to follow threads and participate. By tagging the resources, one mights learn that he is the only one interested in that item, the one person who cared to tag that information. Perhaps he is simply the first one. Perhaps he will become a trend-setter: his acts of tagging will enable others to follow behind, discovering his footsteps.

This works on Flickr, one of the most famous web sites that allows tagging. Someone take a picture. They call it a squared-circle. Another person is struck by that and they create their own squared-circle. Soon, there are many more. Squared-circles become a Flickr trend.

This is the individual perspective. What is happening on the social level with tagging?
4.1.2 Ad-hoc groups or crowds

The basic social formations supported by tagging are more like crowds than true groups. One sees the milling crowds and has some idea about what they are doing, reading, watching, but he does not know these people - they are not part of his network or members of his mailing lists or on-line communities he subscribes to. These are ad-hoc groups brought together by a particular tag or resource.

4.1.3 Four conditions for wisdom of crowds

Starting with Le Bon’s analysis in 1895, psychologists have focused on the negative aspects of crowd behavior. Recently James Suroweicki has refuted this notion. His analysis of four conditions that can lead to wisdom of crowds seems relevant for tagging systems. The four principles are:

1. diversity of opinion: each individual brings its idiosyncratic perspective to bear on the issue.

2. independence of members from one another: people must make independent decisions. This is why mass copying of others’ tags is not a good idea.

3. decentralization: with tagging, power does not reside in a central location

4. good method for aggregating opinions: tag clouds and simple lists seem to work well for this, though better methods are needed.

So far, tagging systems seem relatively free of negative aspects of other types of crowd behavior.
4.1.4 Tagging and Collaborative Filtering

By allowing loose coordination, tagging systems allow social exchange of conceptual information. Collaborative filtering can be likened to a social process: in fact, like-minded individuals share recommendations of books, movies etc.

The process works as follows: one watches a movie, he tells a friend that he liked it. In turn, they recommend a movie to him.

Tagging facilitates a similar but richer information exchange: one comments that a movie is romantic, or a good holiday movie. Everyone who overhears him has access to this metadata about the movie. The social exchange goes beyond collaborative filtering, facilitating transfer of more abstract, conceptual information about the movie.

Tagging enables social coordination that is simultaneously more direct and abstract than collaborative filtering: more abstract since one is exchanging conceptual information, and more direct since there is no algorithm mediating his connection. When one navigates by tags, he is directly connecting with others.

All good social systems need to serve the individual motive. Tagging works because it strikes a balance between the individual and social. It serves the individual motive of remembering, and forms an ad-hoc social groups around it.

4.2 Social Tagging and Social Bookmarking Systems

The concept of shared on-line bookmarks dates back to April 1996 with the launch of itList.com. Within the next three years on-line bookmark services became competitive. The contemporary concepts of social bookmarking and tagging took root
with the launch of the web site one view, in 1999, and Del.icio.us, in September of 2003 [21].

In a social bookmarking system, users store lists of Internet resources, which they find useful. These lists are either accessible to the public or to a specific network, and other people with similar interests can view the links by category, tags, or even randomly. Some allow for privacy on a per-bookmark basis.

They also categorize their resources by the use of informally assigned, user-defined tags. A tag is a relevant keyword or a term assigned to a piece of information, which can be represented by photos, videos or texts, thus describing the item and enabling keyword-based classification of information it is applied to. Tags are usually chosen informally and personally by the creator of the label (who is called tagger) without following a formally defined classification scheme.

Typically, an item will have one or more tags associated with it, as part of some classification software or system. The software will provide links to other items that share that keyword tag, or even to specified collections of tags. This allows for multiple browseable paths through the items which can quickly and easily be altered by the collection’s administrator, with minimal effort and planning.

A tagging example is given to clarify the introduced concepts. A web page hosted on a web server or blog server which supports tagging, might have the tags cats, africa, animal, cheetahs, felis, panthera. A human reader can probably tell the purpose of the page by quickly scanning the list of tags. Typically, the server would display the list of tags as web links leading to an index page which lists all web pages associated to a given tag. This allows a reader to quickly locate all pages which have been associated with the term cheetahs. If the server supports tag searching, a reader would be able to find all pages that use a particular set of tags, such as cheetahs and felis.

If the page’s author wishes to reclassify the page, all that he/she is required to
change is the list of tags. In this case, the author could add the tag *siamese* to such page. All connections between pages are automatically tracked and updated by the server software. There is no need to relocate the page within a complex hierarchy of categories (4.2).

![Diagram of a tagging system](image)

**Figura 4.2.** A model of tagging system

While using tags in such an organizational system is flexible and easy, tagging is not without its drawbacks. Typically there is no information about the meaning or semantics of a tag. For example, the tag *apple* might refer to the fruit, Apple Computer, the Beatles’ music label, or Gwyneth Paltrow’s baby. This lack of semantic distinction in tags can lead to inappropriate connections between items. Additionally, selection of tag terms is highly individualistic. Different people may use drastically different terms to describe the same concept: for example items related to a version of Apple Computer’s operating system might be tagged both *Mac OS X*, *Tiger*, and possibly many other terms. Users of tagging systems must make judgements, based on the number of connections and the choices of tag terms, whether possible connections between items are valid for their interests.

Tag classification, and the concept of connecting sets of tags between web/blog
servers, has lead to the rise of folksonomy classification over the Internet, the concept of social bookmarking, and other forms of social software. Larger-scale folksomonies tend to address some of the problems of tagging, as astute users of tagging system will monitor/search the current use of tag terms within these systems, and tend to use existing tags in order to easily form connections to related items. In this way, evolving folksomonies define a set of tagging conventions through eventual group consensus, rather than by use of a formalized standard.

Although tagging is often promoted as an alternative to organization by a hierarchy of categories, more and more online resources seem to use a hybrid system, where items are organized into broad categories, with finer classification distinctions being made by the use of tags.

4.3 Tagging and Taxonomy

Proponents of collaborative tagging often contrast tagging-based systems from taxonomies.

Familiar taxonomies include the Linnaean system of classifying living things, the Dewey Decimal classification for libraries, and computer file systems for organizing electronic files. In such systems, each animal, book, file and so on, is in one unambiguous category which is in turn within a yet more general one. For example, lions and tigers fall in the genus *panthera* and the domestic cats in the genus *felis*, but *panthera* and *felis* are both part of family *felidae*, of which lions, tigers and domestic cats are all part. Similarly, books on Africa’s geography are in the Dewey Decimal system category 916 and books on South America’s in 918, but both are subsumed by the 900 category, covering all topics in geography.

In contrast, tagging is neither exclusive nor hierarchical and therefore can in some circumstances have an advantage over hierarchical taxonomies. For example,
consider a hypothetical researcher who downloads an article about cat species native to Africa. If the researcher wanted to organize all her downloaded articles in a hierarchy of folders, there are several hypothetical options: each choice reflects a decision about the relative importance of each characteristic. Folder names and levels are in themselves informative, in that, like tags, they describe the information held within them [8].

Despite these limitations, there are several good reasons to impose such a hierarchy. Though there can be too many folders in a hierarchy, an efficiently organized file hierarchy neatly and unambiguously bounds a folder’s contents. Unlike a keyword based search wherein the seeker cannot be sure that a query has returned all relevant items, a folder hierarchy assures the seeker that all the files it contains are in one stable place.

In contrast to a hierarchical file system, a non-exclusive, flat tagging system could, unlike the system described above, identify such an article as being about a great variety of things simultaneously, including *africa* and *cats*, as well as *animals* more generally, and *cheetahs*, more specifically.

Like a Venn diagram, the set of all the items marked *cats* and those marked *africa* would intersect in precisely one way, namely, those documents that are tagged as being about African cats. Even this is not perfect, however. For example, a document tagged only *cheetahs* would not be found in the intersection of *africa* and *cats*, though it arguably ought to: so, the seeker may still need to search multiple locations (4.3).

Looking at it another way, tagging is like filtering; out of all the possible documents (or other items) that are tagged, a filter (i.e. a tag) returns only those items tagged with that keyword. Depending on the implementation and query, a tagging system can, instead of providing the intersection of tags (thus, filtering), provide the union of tags; that is, all the items tagged with *all* of them (4.4). From a user

41
Figura 4.3. A Venn diagram showing the intersection of cats and africa
perspective, navigating a tag system is similar to conducting keyword-based searches; regardless of the implementation, users are providing salient, descriptive terms in order to retrieve a set of applicable items.

![Venn diagram showing the union of cats and africa](image)

**Figura 4.4.** A Venn diagram showing the union of cats and africa

### 4.4 Tagging Systems Problems

Both tagging systems and taxonomies are beset by many problems that exist as a result of the necessarily imperfect, yet natural and evolving process of creating semantic relations between words and their referents. Three of these problems are polysemy, synonymy, and basic level variation.

#### 4.4.1 Polysemy

A polysemous word is one that has many (poly) related senses (semy). For example, a *window* may refer to a hole in the wall, or to the pane of glass that resides within it [19]. In practice, polysemy dilutes query results by returning is related but potentially inapplicable items. Superficially, polysemy is similar to homonymy, where a word has multiple, unrelated meanings. However, homonymy is less a problem
because they can be largely ruled out in a tag-based search through the addition of a related term with which the unwanted homonym would not appear. There are, of course, cases where homonyms are semantically related but not polysemous; for example, searching for employment at Apple may be problematic because of conflicts with the CEO’s surname.

4.4.2 Synonymy

Synonymy, or multiple words having the same or closely related meanings, presents a greater problem for tagging systems because inconsistency among the terms used in tagging can make it very difficult for one to be sure that all the relevant items have been found. It is difficult for a tagger to be consistent in the terms chosen for tags; for example, items about television may be tagged either television or tv. This problem is compounded in a collaborative system, where all taggers either need to widely agree on a convention, or else accept that they must issue multiple or more complex queries to cover may possibilities. Synonymy is a significant problem because it is impossible to know how many items out there one would have liked one’s query to have retrieved, but did not.

Relatedly, plurals and parts of speech and spelling can stymie a tagging system. For example, if tags cat and cats are distinct, then a query for one will not retrieve both, unless the system has the capability to perform such replacements built into it.

4.4.3 Basic Level

Reflecting the cognitive aspect of hierarchy and categorization, the basic level problem is that related terms that describe an item vary along a continuum of specificity ranging from very general to very specific; as discussed above, cat, cheetah and animal are all reasonable ways to describe a particular entity. The problem lies in
the fact that different people may consider terms at different levels of specificity (more general) and subordinate (more specific) interactions with them [25]. For most people, the basic level for felines would be *cat*, rather than *animal* or *siamese* or *persian*. Experiments demonstrate that, when asked to identify dogs and birds, subjects used *dog* and *bird* more than *beagle* or *robin*, and when asked whether an item in a picture is an *X*, subjects responded more quickly when *X* was a basic level [25]. These experiments demonstrate general agreement across subjects.

There is, however, systematic variation across individuals in what constitutes a basic level. Expertise plays a role in defining what level of specificity an individual treats as basic. For example, in the bird and dog experiments, subjects expert in one of the two domains demonstrated basic levels that were at levels of greater specificity than those without domain expertise; a dog expert might consider *beagle* a basic level where a bird expert might have *dog* and a bird expert *robin* where a dog expert has *bird* [25].

The underlying factor behind this variation may be that basic levels vary in specificity to the degree that such specificity makes a difference in the lives of the individual. A dog expert has not only the skill but, also the need to differentiate beagles from poodles, for example. Like variation in expertise, variations in other social or cultural categories likely yield variations in basic levels.

For the purposes of tagging systems, however, conflicting basic levels can prove disastrous, as documents tagged *perl* and *javascript* may be too specific for some users, while a document tagged *programming* may be too general for others.

Tagging is fundamentally about sense-making. Sense-making is a process in which information is categorized and labeled and, critically, through which meaning emerges [29]. Recall that basic levels are related to the way in which humans interact with the items at those levels [25]; when one interacts with the outside world, one makes sense of the things one encounters by categorizing them and ascribing
meaning to them. However, in practice, categories are often not well defined and their boundaries exhibits vagueness [9]. Items often lie between categories or equally well in multiple categories. The lines one ultimately draws for oneself reflect one’s own experiences, daily practices, needs and concerns.

Sense-making is also influenced by social factors [29]. Because many experiences are shared with others and may be nearly universal within a culture or community, similar ways of organizing and sense-making do result; after all, societies are able to collectively organize knowledge and coordinate action. Additionally, collective sense-making is subject to conflict between the participating actors, where different opinions and perspectives can clash and power struggles to determine the terms of the debate can ensue [29].

Collective tagging, then, has the potential to exacerbate the problems associated with the fuzziness of linguistic and cognitive boundaries. As all taggers’ contributions collectively produce a larger classification system, that system consists of idiosyncratically personal categories as well as those that are widely agreed upon. However, there is also opportunity to learn from one another through sharing and organizing information. Actually this thesis work is mainly focused on the problems related to words, in a synonyms and structural sense. In fact, using two different approaches, our study try to give a solution to the words meaning-disambiguation problem. On the other hand, if future works will solve also the other two main problems (polysemy and basic level), then searching engines could effectively become concepts-based instead of keyword-based and the creation of ontologies based on tags will get reality.
4.5 Social Tagging Tools

Differently from the simple bookmark managers built into many browsers, the new breed of link managers is a server-side web application that enables links to be tagged for easy retrieval and is increasingly being opened up to manage public rather than (or sometimes in addition to) private sets of links.

Social bookmarking sites are in many ways the ultimate folksonomies, categorizing web content by an open process of tagging. The result is a road map of the web, based on the recommendations of people who have traveled it extensively.

In many ways, the information gathered by social bookmarking services is like the wild, young sibling of the web directories. But now, every web surfer is a potential editor, licensed to categorize and recommend all kinds of web sites.

The following elements are usually present in varying degrees:

- A decent amount of users is necessary. Otherwise the whole social aspect looses relevance.

- Personal user accounts (groups sometimes provided)

- Mechanism for entering links, titles and descriptions

- The service has to have quite extensive documentation, FAQ and/or help pages, to guide users.

- Browser bookmark-lights to facilitate entry: at the time of writing allows to post to a dozen social bookmarking tools with one click

- Classification by 'open' or 'free' tagging

- Search by tag or user (Boolean combinations sometimes allowed)

- Querying of links based on popularity, users, tags, etc.
More advanced users will want good import and export possibilities. Import is necessary for bringing old bookmarks from your browser or from another on-line bookmarking service, export is important for backup copies and to ensure that your bookmarks can go with you to another service if you decide to switch.

A good bookmarking service should have feeds (RSS and/or Atom) of tags or groups of tags.

Extensions such as browser plug-ins

Have blogs so that one can be kept up to date about new features, bugs, bug fixes and more.

There are many open bookmarking tools available, a number of which fall loosely within the category of social bookmarking tools, so that they include some kind of a tagging or rating element.

4.1 presents an overview of some of the tools that we have selected to review and provides a comparison of their public visibility, including history and usage, and business models. 4.2, by contrast, shows the feature sets for these same tools.

What is striking from 4.1 is that all these tools are of very recent public origin (almost all from 2004) although they may have had a longer private gestation. Also the tools we analyze here are equally split between academic, commercial and private sponsorships.
<table>
<thead>
<tr>
<th>Tool</th>
<th>Sponsor</th>
<th>Launch</th>
<th>Users</th>
<th>Links</th>
<th>Tags *</th>
<th>Business Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connotea</td>
<td>Nature Publishing Group (UK)</td>
<td>Dec.'04</td>
<td>&gt;2k</td>
<td>&gt;4k</td>
<td>&gt;3k</td>
<td>Supported by NPG as potential driver to other products</td>
</tr>
<tr>
<td>Del.icio.us</td>
<td>Privately Managed (US)</td>
<td>Dec.'03</td>
<td>50k</td>
<td>1m</td>
<td>2m</td>
<td>Advertising being considered</td>
</tr>
<tr>
<td>Flickr</td>
<td>Ludicorp (Canada)</td>
<td>Feb.'04</td>
<td>245k</td>
<td>3.5m</td>
<td>-</td>
<td>Annual premium account($60), Google Ads</td>
</tr>
</tbody>
</table>

Tabella 4.1. Reviewed Social Bookmarking Tools - Comparison

* Note that this column is intended to reflect the number of 'unique' tags. It is not always clear whether the numbers recorded here are for unique tags or total number of tags. For example, the number of tags quoted for Del.icio.us is 2m, which may be rather high in view of the number of tags exceeding the number of links and also that the entire English vocabulary may only be around some 300k words, although note that tags are not limited either to 'words' or to a particular language such as English.

From 4.2 we can see that the academically inclined bookmarking tools offer RSS 1.0 with Connotea providing rich Dublin Core and PRISM. It is also encouraging to see that one of the most popular bookmarking services, Del.icio.us, is also offering RSS 1.0 with Dublin Core.
Two of the best known web sites that allow social tagging are *Del.icio.us* and *Flickr*. As visible from 4.2, neither of them provide the source code of the platform, but it is useful to analyze their characteristics to locate an open source tool that offers similar functionalities.

### 4.5.1 Closed source platforms

**Del.icio.us**

Del.icio.us is by far the most popular social bookmarking tool out there. It is not the most powerful, though. There were some fifty thousand users on April, 2005. This is very much a personal, out-of-hours effort by Joshua Schachter without any immediate intention of developing a business model. Developed initially as a simple web page listing links with annotations, Schachter then decided to make these available on a web server so that friends and others could also make use of these bookmarks. From there, it was but a small step to hosting others’ lists, and so the social bookmarking tool Del.icio.us was born.

In this tagging system, the tagged objects are bookmarks related to generic web pages. Each registered user can allocate one or more tags to whichever bookmark and has the possibility to rename, eliminate, add and aggregate tags. Moreover each user has its own tags list that can be made visible to other users to allow tag reuse. However the tools for sharing are not very powerful: you can add people to your network and share links with them, and that’s all about it. The effect of sharing can
be great, though, because of the size of the community. Each time that a user tags a new bookmark, it is saved in its personal profile within the following informations:

- title of the web page
- short description of the web page
- personal notes
- associated tags.

On Del.icio.us you can view your bookmarks as a tag list or a tag cloud. Tags can be sorted alphabetically or by frequency both at personal profile level and at global level.

Behind the Help link you find tips for advanced searching (by user, by tag, by both user and tag) and navigation and some unofficial tools that can improve Del.icio.us functionalities for more demanding users. Among the advanced features are network badges, link rolls, and tag rolls that help bloggers and webmasters to display their Del.icio.us activities on their own blogs and sites.

There are no very powerful tools for sorting or editing bookmarks. Paradoxically, this might be a reason for Del.icio.us’ popularity. It offers just what the average web surfer needs.

**Flickr**

Flickr is used for managing images (specifically digital photos) rather than links. Co-founded by Stewart Butterfield and Caterina Fake, it is now run as a commercial operation under the Ludicorp brand.

Flickr has the widest user base of all these tools. This, perhaps, is not so surprising because it is built around photos, digital cameras and camera phones being prevalent these days, and there are very strong emotional drivers for sharing images.
Flickr hosts users’ photos on its own servers so that they are retrievable (in various size samplings) from anywhere over the Web.

It provides free user accounts, albeit bandwidth-capped for uploading, and premium paid-for accounts that limit the restrictions that are placed upon users.

The base principles are the same of Del.icio.us but there are some differences: the images can be shared with the general public, only with Flickr’s registered users or with a restricted number of people (e.g. friends, relatives, contacts). A specific user can allow other selected users to modify the tags associated to its own images collection (this is not possible in Del.icio.us). Furthermore it is possible to associate Creative Commons licenses to images, that specify the ways in which a given image can be (re)utilized from a user that downloads it.

Photos can be aggregated into photo-sets and can be made public or private as desired.

### 4.5.2 Open source platforms

Thanks to the increasing success of social tagging web sites, a lot of enterprises have looked for, shared and tested source code that provide functionalities similar to the ones offered by the tools mentioned above.

Some of these platforms have been analyzed in these sections with particular attention to the supplied functionalities. In particular we have looked for tools that allow to start from a point as closed as possible to the current state of art, with consequent benefits for future research developments.

**Connotea**

Connotea project was born to share information between scientists and clinicians. A public web portal has been built, based on Connotea, to insert, edit and search tags and information: saving references in Conntoea is quick and easy. One can
do it by saving a link to a web page for the reference, whether the PubMed entry, the publisher’s PDF, or even an Amazon product page for a book. Connotea will, wherever possible, recognise the reference and automatically add in the bibliographic information for the user. Every time one wants to save a reference to his/her library, simply clicks the Connotea browser button, and a form to add the reference to his/her library will pop up. Connotea will automatically fill in the title and URL for the page, and also add in bibliographic information like author and journal names for a growing list of websites. All the user needs to do is add some tags, a description or comment if he/she wishes, and clicks a button to add the reference to his/her library. The pop-up will disappear, leaving the person back where he/she started, with minimal interruption to his/her work.

At the time of writing, all bookmarks posted to Connotea are visible to all registered users and visitors. This takes the concept of sharing to a new level, but also brings new opportunities. It is acceptable that some users want to keep some of what they are reading as personal, therefore private bookmarks are soon added as a feature, though the default is to make the bookmark public. The main benefits of openness come not just from the ease with which it allows explicit sharing with friends and colleagues, but from many users storing their bookmarks in the same place. This allows Connotea to automatically discover and present connections between users. For example, if someone else has bookmarked the same things as one specific user, that person’s library will be a good candidate for a place to find interesting new content. In addition, shared lists allow more sophisticated collaborative filtering algorithms to make recommendations of the form people who bookmarked this also bookmarked....

In Connotea, each one assigns tags to his/her own references. These can be anything he likes, and he can use as many as he like, so there’s no more need to navigate complicated hierarchies of folders and categories. Connotea shows to the
user all the tags he ever used, so it’s easy to get back to a reference once he has saved it.

Each user can explore other Connotea users’ libraries just as easily as he can navigate his own. By saving his references on Connotea, he’s connecting his reading to that of other Connotea users who are working on the same things. One can then click on related tags and related users to discover new articles and links.

The supporting code offers the minimal functionalities like the creation, modification, cancellation of tagged objects; the request of bookmarks, tagged objects or tags filtered by user, tag, tagging date, URI, etc. It seems not to provide advanced functionalities like correlated objects search, statistics collections or similar.

The available implementations are in Perl, Ruby and Python, that are all not particularly diffused and well known programming languages.

**Del.rio.us**

Born from an idea of Steve Mallet, Del.rio.us project is a Del.icio.us clone as can be easily noticed by the name of the web site. Del.rio.us follows on the concept of tagging content and sharing information and links by allowing users to subscribe to the RSS of each tag. The difference between Del.rio.us and Del.icio.us is that:

- you can add notes,

- you can use it as a blog.

In particular, each user can post website addresses that find interesting. Furthermore, he can also include a brief description, and longer detailed notes (these can also be blog posts).

The source code is divided in two main parts: the first one, *Rubric*, provides the libraries with the functionalities related to the users, the tags and the database administration; the second one has an higher level purpose and it is mainly composed
by template of web pages that exploit *Rubric* to offer interaction to the users. This last part seems no longer used (the author himself suggests the migration to Simpy, another commercial web site based on tagging). The API of *Rubric* seems to be quite well structured and documented, though they do not provide particularly advanced functionalities for the search and the management of the tags.

The source code is written in Perl.

**FreeTAG**

FreeTAG is an easy tagging and folksonomy-enabled plug-in for use with MySQL-PHP applications. It allows to create tags on existing database schemes, and access and manage tags through a robust API.

In particular, in addition to the basic functionalities, some advanced functions are available. Examples of these are:

- the search of the recent tagged objects,
- the search of similar (to a given one) tags,
- the search of similar (to a given one) objects.

The platform organization is sufficient simple, clear and contemporary general: it seems suitable for integration with other platforms furthermore to provide a valid starting point for more search development, such as tag clustering and semantics.

**Serendipity**

Serendipity is a PHP-powered web blog application which gives the user an easy way to maintain an on-line diary, weblog or even a complete homepage with a social tagging approach. While the default package is designed for the casual blogger, Serendipity offers a flexible, expandable and easy-to-use framework with the power for professional applications.
Casual users appreciate the way Serendipity’s sophisticated plug-in architecture allows to easily modify both the appearance of the blog and its features. One can install more than 120 plugins with just one click, instantly enhancing blog’s functionality without requiring code editing.

Likewise, one click installs any of more than 40 official templates, facilitating appearance customization. Furthermore, Serendipity’s SPARTACUS plug-in automatically checks the central repository for upgrades and new functionalities.

Advanced users value Serendipity’s Smarty templates for combining simplicity with well-documented web standards. It makes minor modifications trivial, but provides the power to unleash creativity and customizations. Serendipity’s outstanding support gives the confidence to be adventurous, too.

Programmers and other technical users recommend Serendipity for its fast, stable and clean PHP code. While beginners can learn from Serendipity, advanced programmers can easily make complex modifications. Serendipity is programmed in PHP, long recognized for its ideal blend of power, simplicity, and speed. Serendipity’s BSD licensing ensures that programmers around the world can learn from it and improve it.

Users of other blogging/CMS applications are already switching to Serendipity, thanks to its easy customization and outstanding support. Corporate users are taking advantage of Serendipity’s unparalleled flexibility to set up fast, simple CMS sites.

Serendipity’s basic features include something for everybody, from the personal blogger to the professional corporate web designer:

- WYSIWYG and HTML editing,
- Built-in, powerful media database,
- Multiple authors, configurable permission/user-group system,
- Threaded comments, nested categories, post to multiple categories,
- Multiple languages (internationalization),
- On-line plug-in and template repository for easy plug-and-play installation,
- Cool plug-ins: category-based sub-blogs, podcasting, RSS planet/aggregator, static pages,
- Robust spam blocking,
- One-click upgrading from any version,
- Can be embedded into one existing web pages,
- Standards-compliant templating through Smarty, remote blogging via XML-RPC,
- BSD-style licensing,
- Multiple Database support (SQLite, PostgreSQL, MySQL, MySQLi),
- Shared installations can power multiple blogs from just one codebase,
- Native import from earlier blog applications (WordPress, Textpattern, Moveable Type, bblog, etc.).

The source code is prevalently written in PHP and is actively supported, but it seems not to offer easy hints for the integration with new contexts.

However, the effective tagging functionalities are not clear since the related documentation is incomplete and it is necessary a deep source coding study.

This project can surely supply useful cues but it is too articulate and inadequate to be used as simple starting point for the integration and development of our research purposes.
Pligg

Pligg is an Open source Social Networking Content Management System (CMS) Combining social bookmarking, blogging, and syndication and a democratic editorial system enables users to collaboratively submit and vote articles. It was influenced by the extremely popular English technology site digg, where when a user submits a news article it is placed in the pligged area until it gains sufficient votes to be promoted to the main page.

Pligg is unique compared to most other content management systems because of it’s flexibility. A web designer can do pretty much anything with Pligg because the software was designed to be used in as many ways as possible. Not only can a person with very little knowledge of PHP and MySQL install it, but they can modify and administer it with relatively little difficulty. For those who have a greater understanding of web languages, Pligg can act as the first step in a highly customized personal content management system.

Pligg’s dynamic structure allows to quickly and effortlessly install and customize blog-based sites. Pligg acts as a framework for many successful sites who have taken the code offered by Pligg and added their own touch to it.

Also this project, similarly to Serendipity is quite rich and articulated and provide templates for the creation of a web site. Furthermore it keeps trace of some statistics related to the frequency and goodness of the tags inserted by each user.

The available documentation is minimal and it have to be studied with attention to identify a good integration strategy.

The source code is written in PHP.

Scuttle

Scuttle is an open source social bookmarking application that offers functionalities similar to Del.icio.us ones without the shortcomings. It is available under the GNU
General Public License (GPL), and requires PHP and MySQL.

With Scuttle there is no need to start from scratch: the application allows to import a user local browser and Del.icio.us bookmarks. Scuttle allows to specify one of three types of security settings for each bookmark in one’s personal collection: public, private, or shared with watch list. Public bookmarks are available to anyone, including guests that are not registered with Scuttle. If one sets a bookmark status to shared with watch list, it will be available to registered users, but not to unauthorized users. Finally, private bookmarks are available only to a specific user. Each security setting has its own color code, which makes it easy to identify a bookmark’s status.

Actually, Scuttle not only offers the ability to import bookmarks, it also supports most of the Del.icio.us API. This means that virtually any Del.icio.us tool can work with Scuttle, though it must be modified first.

Like Del.icio.us, Scuttle exports RSS feeds, so one can keep an eye on other users’ public bookmarks. However, Scuttle takes the idea of sharing bookmarks even further. Using the Watchlist feature, users can view each other’s public bookmark listings, and add interesting bookmarks in the watch list to their own bookmark collection by simply clicking on the Copy link.

Scuttle has several other features not available in Del.icio.us. All tags in Scuttle are presented as a tag cloud, making it easier to locate the most popular tags. The tags can be sorted either alphabetically or by popularity. The same goes for bookmarks: one can sort them by date, title, or URL. Scuttle also supports the related tags feature that allows to view bookmarks with multiple tags. And when one adds a bookmark containing a link to an audio file in MP3 or Ogg format, Scuttle places a tiny player icon next to it, and you can play the file without leaving Scuttle.

Very important for international user is the multilingual nature of Scuttle: all text strings in the application are stored in gettext (.po) catalogs, which makes
it easy to localize software to other languages using tools like poEdit. Several translations are already available on Scuttle’s wiki.

The source code is mainly written in PHP and supports templates and some other advanced functionalities like watch lists administration and user authorization, in addition to advanced user interfaces written in conformity with the recent Ajax technologies. It is surely a valid reference point but the documentation is minimal and eventual extensions or modifications require a careful code analysis since it is certainly less manageable with respect to simpler projects.

4.5.3 Final Consideration

After considering the available documentation, the complexity and the completeness, the generality and the integration simplicity of the open source analyzed platforms, we have decided to use FreeTAG as starting point for our study. The chosen solution seems to be the simpler and generic, suitable for extensions and adaptations in addition to the essential functionalities. Perhaps other social tagging platforms could be used for future integration and improvements: for example we think about multi-languages administration, necessity of template and plug-in (like Serendipity), statistics based on users’ reliability (like Pligg), watch-list, advanced interfaces and users’ authorizations (like Scuttle).
Capitolo 5

Results Analysis

In this Chapter the final results obtained from both the statistical analysis and the experimental work are presented and discussed.

5.1 Statistical Analysis Results

5.1.1 Users activity and Tag Quantity

In our dataset, obtained from the Connotea website, there is a very strong relationship between the age of the user’s account (i.e. the time since they created the account) and the number of days on which users create at least one bookmark. Visibles, many users who have no tags in their list. Usually these users are the most recent ones, e.g. the ones that have just registered but not yet inserted any tag (5.1).
Figura 5.1. The number of tags in each user’s tag list, in decreasing order
5.2 shows a detail of the previous plot that highlights how some users use Connotea very frequently, while others much less frequently.

Figura 5.2. Detail of the number of tags in each user’s tag list, in decreasing order
More interestingly, there is a weak relationship between the number of bookmarks a user has created, and the number of tags he/she uses for those bookmarks. The relationship is weak at the low end of the scale, there are users with fewer than 30 bookmarks, and even weaker at the upper end, where some users have more than 200 bookmarks. Furthermore, as evidenced by 5.3, some users have comparatively large sets of tags, while other users have comparatively small sets. In other words, it comes out that as the number of bookmarks increases, the number of used tags also increases but at very different rates. It is possible to notice how the growth rate of tags inserted by user #6437 increases very rapidly with respect the one of user #20665: the second one often uses the same keywords while bookmarking.

Figura 5.3. Two extreme users’ (#6437, #20665) tag growth. As they add more bookmarks, the number of tags they use increases, but at very different rates.
Users’ tag lists grow over time. Tags may exhibit very different growth rates, however, reflecting that interests develop and change over time. 5.4 and 5.5 show how the use of each tag increases as each user adds more bookmarks over time.

5.4 depicts the growth rate of user #6437. All the considered tags have a linear and constant growth till a certain point; then, the increase becomes steady. Thus, the interests of such user vary so radically over time that he completely forsakes the use of the keyword from a moment on.

![Growth Rates of Three Tags](image)

**Figura 5.4.** Growth rate of three selected tags for user #6437
5.5 represents the tagging behavior of user #20665. A very strange trend can be noticed. In fact, for a certain period of time, the number of considered tags linearly grows, while from the 200\textsuperscript{th} bookmark on, tags growth become again null: there is a tag that increases rapidly, reflecting a newfound interest or a change in tagging practice, whereas some other remains practically constant, reflecting a probable decay of such interest.

Figura 5.5. Growth rate of five selected tags for user #20665
There is a last thing that must be underlined: user #6437 (5.4) did not use tag #80861 until approximately the 250th bookmark. If tag #80861 indeed constitutes a new distinction among the kinds of items this user bookmarks, it would be very difficult to reconsider each of the earlier 250 bookmarks to decide whether to add tag #80861 to them. Further, if in future this user needs to filter his bookmarks by such tag id, then no bookmark before the 250th will be retrieved, compromising the practical usefulness of the tag.

5.4 and 5.5 show how the interests of the users evolving and growing. Next, the attention is turned to discover what roles tags play in bookmarks.

5.1.2 Resources

In this part of the study the addition of bookmarks to URLs over time is analyzed.

Usually URLs receive most of their bookmarks very quickly, while the rate of new bookmarks decreases over time.

As well our analysis exploits this fact: typically, URLs reach their popularity very rapidly, in the first few days.
Figura 5.6. The addition of bookmarks to URL #206083 over time.
Figura 5.7. The addition of bookmarks to two URL #206959 over time.

Figura 5.8. The addition of bookmarks to URL #234283 over time.
Only in one case (5.9) the URL takes a longer time (8 days) to become popular: since that day users have not noticed such resource and then they start to bookmark it making its popularity growing up.

Figura 5.9. The addition of bookmarks to URL #235462 over time.
Furthermore, 5.6, 5.7, 5.8 and 5.9, show a maximum number of received bookmark which is quite small. This is a very strange behavior justifiable considering that the bookmarked resources are usually related to very specific subjects; therefore only few people are interested to them.

### 5.1.3 Kinds of Tag

Some tags are used by many people, while other tags are used by fewer people. Additionally, those tags that are generally meaningful will likely be used by many taggers, while tags with personal or specialized meaning will likely be used by fewer users.

5.10 shows how tags used to characterize resource #106803 are the more frequent ones: in fact such resource is more recent with respect resource #241990, it has been inserted afterwards. So, users that tag the newer resource can use keywords already used by other in the past. Typically these tags will be the most general ones in a meaning sense, so, the ones that glue different users interests.

In 5.10 is displayed the median rank computed as the frequency of the tags orders in a decreasing way: specifically, more frequent tags have smaller values of median rank.

5.11 shows that there are only few resources which have been tagged with several tags but many have received few different keywords. This result point out how tags are spread and make actual and even more necessary the study of possible aggregation of tags as a way to link resources which are in some way correlated from a conceptual perspective.

Most of the results obtained from the statistical analysis of the Connotea dataset can be compared with the ones achieved by [5], since our analysis follows the baseline of the mentioned work. Particularly, the results obtained from the two works are well comparable under some aspects (i.e., the number of tags in each user’s tag list.
Figure 5.10. As tags’ order in a bookmark (horizontal) increases, its rank in the list of tags (vertical) decreases. This pattern is shown here for several URLs (#106803, #241990, #51626, #106726, #106753).

Figure 5.11. The number of distinct tags per resource, in decreasing order.
or the behavior of two extreme users’ tags growth, but very different under some others (i.e., the growth rate of some selected tags or the addition of bookmarks to the URLs over time) due to the intrinsic different nature of the considered data.

In general, collaborative tagging users exhibit a great variety in their sets of tags: some users have many tags, and other have few. Tags themselves vary in frequency of use, as well as in what they describe.

The prevalence of tagging with a very large number of tags and according to information intrinsic to the tagger demonstrates, due to the fact that minority opinions can coexist alongside extremely popular ones, that a significant amount of tagging, if not all, is done for personal use rather than public benefit. Nevertheless, even information tagged for personal use can benefit other users. Information tagged by others is only useful to the extent that the users in question make sense of the content in the same way, so as to overlap in their classification choices.

The stable, consensus choices that emerge may be used on a large scale to describe and organize how web documents relate to one another. Starting from this point it is possible to suppose that tagged resources may be valuable in aggregate as well as individually.

5.2 Experimental Work Results

5.2.1 Tags

The first part of the experimental work is devoted to evaluate the actual correlation between a tag and the first resources associated to it.

5.1 shows the recall obtained using the three different techniques, which, as expected, return an increasing amount of similar resources associated to the given tag. Specifically the Stemming and WordNet approaches give, as an average, 40 and 269 more resources than the FreeTAG approach respectively.
## 5 – Results Analysis

Tabella 5.1. Number of resources correlated to the given tag

<table>
<thead>
<tr>
<th>Tag</th>
<th>Number of Returned Resources</th>
<th>FreeTAG</th>
<th>Stemming</th>
<th>WordNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bacteria</td>
<td></td>
<td>420</td>
<td>420</td>
<td>468</td>
</tr>
<tr>
<td>Clostridium</td>
<td></td>
<td>516</td>
<td>516</td>
<td>987</td>
</tr>
<tr>
<td>DNA</td>
<td></td>
<td>452</td>
<td>455</td>
<td>1023</td>
</tr>
<tr>
<td>Genetics</td>
<td></td>
<td>1187</td>
<td>1439</td>
<td>1617</td>
</tr>
<tr>
<td>H5N1</td>
<td></td>
<td>937</td>
<td>938</td>
<td>958</td>
</tr>
<tr>
<td>HIV</td>
<td></td>
<td>1507</td>
<td>1507</td>
<td>1845</td>
</tr>
</tbody>
</table>
The evaluation is made on these 7 tags extracted by a FreeTAG function (get_most_popular_tags.php) as the most popular: a PHP array with tags ordered by decreasing popularity is returned. For each of the 7 considered tags, the first 10 associated resources are taken into account. This operation is made for all the three adopted approaches (FreeTAG, Stemming and WordNet). The obtained resources have been manually compared in meaning to the initial tags as follows: if the tag is mentioned in the title or body of the document, such resource is labeled with yes, to indicate that the resource is clearly related to the tag; if neither the title nor the body of the resource have an explicit relationship with the given tag, but the context allows to understand the relation resource-tag, the more or less label is used; in all other situations the no label is used. It should be noted that all the involved resources have been manually retrieved and matched against the tag, so requiring a considerable amount of time. In some cases the resources are not available on the web, and are therefore labeled as not found.
FreeTAG

As showing in 5.12, for all the considered tags most of the resources returned by FreeTAG (y axis in the figure) are actually correlated to the initial tag (x axis).

Figura 5.12. FreeTAG Results for tag-resource correlations
Stemming

Slightly better results than the ones returned by FreeTAG are obtained using the stemmed tags as shown in 5.13. In this case the approach described in Section 5.4.2 has been used to retrieve the resources, given a tag.

![STEMMING RESULTS](image)

**Figura 5.13.** Stemming’s Results

WordNet

The results obtained with the WordNet approach described in Section 5.4.3, are, in general, better than the ones just presented. Particularly, a 100% of resource-tag affinity is obtained for *HIV, Genetics* and *Clostridium*. Only in two cases (*DNA* and *Bacteria*) worse results are obtained.
Figura 5.14. WordNet Results
Results Comparison

To better compare the results for the three approaches in 5.15, 5.16 and 5.17 are plotted respectively the number of resources labeled with *yes, more or less* and *no*.

![Yes Comparison Diagram](image)

**Figura 5.15. Yes Comparison**
Figura 5.16. More Or Less Comparison
It can be easily noted how in 3 out of the 7 considered cases WordNet provides the best results. In all the remaining cases FreeTAG and Stemming approaches provide similar results except for the *Genetics* tag for which Stemming gives better results.

Furthermore is it important to notice that the resource-tag correlation is strictly dependent on the generality of the tags: the more specific is the tag (e.g. *HIV*) the better are the obtained results. Viceversa the more general is the tag general (e.g. *Bacteria*), the less are the relevant results.

In 5.2 are presented the percentage of the global performances.

### 5.2.2 Resources

After the results for tag-resource correlation, we discuss here the results for resource-resource correlations, which would help the users to browse resources by similarity.
<table>
<thead>
<tr>
<th>Label</th>
<th>Average Resources Per Tag</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FreeTAG</td>
</tr>
<tr>
<td>yes</td>
<td>74%</td>
</tr>
<tr>
<td>more or less</td>
<td>19%</td>
</tr>
<tr>
<td>no</td>
<td>1%</td>
</tr>
<tr>
<td>not found</td>
<td>6%</td>
</tr>
</tbody>
</table>

Tabella 5.2. Average percentage of tags effectively correlated to the resources
In this case the goal is to aggregate different resources basing on the associated tag or set of tags. The three different approaches presented in Section 5.4 are used and, for each of them, 17 have been chosen for testing purposes. In particular, for each resource, the first 10 resources having some tag in common are considered and analyzed qualitatively in terms of topic similarity.

Additionally, to facilitate plots reading, it is necessary to clarify what are the adopted evaluation criteria: if the returned resources, considered as similar, have exactly the same subject matter, then they are labeled with yes; if they are only partially related to the given resource topic, then they are labeled with more or less, and if they represent completely different subjects then they are labeled with no.

Finally, it is important to highlight that following results are used to evaluate the goodness and the precision of the approaches (Section 5.4). Moreover, 5.3 shows the recall obtained using the three different techniques, which, as expected, return an increasing amount of similar resources. In particular, from 5.3 it is possible to infer that the Stemming and WordNet approaches give as an average 676 and 4835 more resources than the ones obtained using FreeTAG technique.
### Results Analysis

<table>
<thead>
<tr>
<th>Resource ID</th>
<th>Number of Similar Returned Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FreeTAG</td>
</tr>
<tr>
<td>16664</td>
<td>595</td>
</tr>
<tr>
<td>23535</td>
<td>2187</td>
</tr>
<tr>
<td>50553</td>
<td>1204</td>
</tr>
<tr>
<td>52162</td>
<td>450</td>
</tr>
<tr>
<td>60696</td>
<td>298</td>
</tr>
<tr>
<td>86999</td>
<td>244</td>
</tr>
<tr>
<td>106005</td>
<td>273</td>
</tr>
<tr>
<td>106390</td>
<td>212</td>
</tr>
<tr>
<td>165926</td>
<td>347</td>
</tr>
<tr>
<td>169591</td>
<td>300</td>
</tr>
<tr>
<td>201064</td>
<td>309</td>
</tr>
<tr>
<td>210157</td>
<td>849</td>
</tr>
<tr>
<td>214161</td>
<td>1664</td>
</tr>
<tr>
<td>218086</td>
<td>319</td>
</tr>
<tr>
<td>239207</td>
<td>1522</td>
</tr>
<tr>
<td>240918</td>
<td>1182</td>
</tr>
<tr>
<td>242582</td>
<td>2257</td>
</tr>
</tbody>
</table>

Tabella 5.3. Number of resources correlated to the given one
FreeTAG

As shown in 5.18, the results obtained with this kind of analysis are generally good. In fact for 14 resources out of the 17 considered ones the majority of the correlated resources are effectively similar in topic, thus, they have been labeled with *yes* or *more or less*. For 8 tested resources the correlated resources labeled with *yes* are more than 50%.

Figura 5.18. FreeTAG Results
Stemming

Slightly worse results, in term of precision, are obtained using this approach. In fact, only 6 resources have in the results set a number of *yes* bigger than the 50%. However, considering also the resources only partially similar (labeled with *more or less*), the obtained results are comparable with the ones obtained using FreeTAG approach.

Since the obtained results are not as good as expected, a further statistical analysis is performed: for each stemmed tag is found the number of distinct tags associated to it ordered in a decreasing way (5.20).

As visible from 5.20 there are a lot of stemmed tags associated to only one tag: this suggests that Stemmed approach is not good in all these cases.

Solutions could be to use another stemming algorithm, based on other principles, or to adopt a different approach to aggregate different keywords. This last one is the solution chosen in our study.

Figura 5.19. Stemming Result

Since the obtained results are not as good as expected, a further statistical analysis is performed: for each stemmed tag is found the number of distinct tags associated to it ordered in a decreasing way (5.20).
Figura 5.20. Number of distinct tags associated to each stemmed tag, in a decreasing order
WordNet

In this case a considerable improvement in performances was expected. Contrary to all expectations, the provided results are generally as good as the one supplied by the FreeTAG approach.

![WordNet Results](image)

Figura 5.21. WordNet Results

Furthermore an awkward behavior is pointed out: a certain number of resources are labeled as *not found*. This complicates result evaluation because, in such cases, it is not possible to assess whether the returned resources are effectively similar to the input one, so impacting the result comparison. The reasons of the *not found* outcome are mainly two:

- the URL of the resource is incorrect or incomplete
- the supplied web page is no longer working.
Results Comparison

5.22, 5.23 and 5.24 allow to evaluate whether an approach is better with respect to the other ones. Considering the number of *yes* that a resource receives, and, with an equal number of *yes*, considering also the number of *more or less* labels, we can conclude that:

- in three cases FreeTAG provides the best results (resources #23535, #106390, #210157)
- in seven cases the Stemming yields the best performances (resources #52162, #106005, #165926, #201064, #218086, #239207, #242582)
- the WordNet approach is the best in the remaining seven cases (resources #16664, #50553, #60696, #86999, #169591, #214161, #240918).

Fig. 5.22. Yes Comparison

---

89
5 - Results Analysis

Figura 5.23. More or Less Comparison

Figura 5.24. No Comparison
5.4 shows the obtained performances.

<table>
<thead>
<tr>
<th>Label</th>
<th>Average Correlated Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FreeTAG</td>
</tr>
<tr>
<td>yes</td>
<td>39%</td>
</tr>
<tr>
<td>more or less</td>
<td>31%</td>
</tr>
<tr>
<td>no</td>
<td>26%</td>
</tr>
<tr>
<td>not found</td>
<td>4%</td>
</tr>
</tbody>
</table>

Tabella 5.4. Average percentage of resources effectively correlated to the given one

The obtained results shown in 5.4, seem to indicate that the Stemming and WordNet approaches are generally less precise than the base-line approach. However, the returned resource are not ranked well and it is reasonable to expect that many retrieved resources are actually relevant, yet not presented between the just considered. To better understand this situation another experiment is presented. In this step are chosen the resources for which the WordNet approach returns 9 correlated resources. This should be a way to by-pass the ranking problem in order to see if the obtained results are good or not (as it is impossible to manually check thousands resources).

Also in this step the three different approaches presented in Section 5.4 are used and, for each of them, 17 have been chosen for testing purposes.

The labels used in the plots have the same meaning as before and in 5.5 it is possible to see the recall obtained using these three different techniques, which, as expected, return an increasing amount of similar resources for Stemming and WordNet respectively. In particular, from 5.5 it is possible to infer that Stemming approach gives in average 2 more resources than the ones obtained using the FreeTAG technique while WordNet returns (as an average) 3 more resources (still than FreeTAG).
### Table 5.5. Number of resources correlated to the given one

<table>
<thead>
<tr>
<th>Resource ID</th>
<th>Number of Similar Returned Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FreeTAG</td>
</tr>
<tr>
<td>14470</td>
<td>6</td>
</tr>
<tr>
<td>19680</td>
<td>3</td>
</tr>
<tr>
<td>31696</td>
<td>6</td>
</tr>
<tr>
<td>33498</td>
<td>7</td>
</tr>
<tr>
<td>43088</td>
<td>5</td>
</tr>
<tr>
<td>72726</td>
<td>7</td>
</tr>
<tr>
<td>76557</td>
<td>6</td>
</tr>
<tr>
<td>111634</td>
<td>3</td>
</tr>
<tr>
<td>116800</td>
<td>6</td>
</tr>
<tr>
<td>124613</td>
<td>6</td>
</tr>
<tr>
<td>151865</td>
<td>8</td>
</tr>
<tr>
<td>166290</td>
<td>6</td>
</tr>
<tr>
<td>169614</td>
<td>3</td>
</tr>
<tr>
<td>186019</td>
<td>8</td>
</tr>
<tr>
<td>205255</td>
<td>8</td>
</tr>
<tr>
<td>214753</td>
<td>7</td>
</tr>
<tr>
<td>240311</td>
<td>8</td>
</tr>
</tbody>
</table>
FreeTAG

As shown in 5.25, the results obtained with this kind of analysis are generally good. In fact for 9 resources out of the 17 considered ones the majority of the correlated resources are effectively similar in topic, thus, they have been labeled with yes or more or less. Unfortunately, many resources could not be retrieved and checked (resources labeled as not found).

![Figure 5.25. FreeTAG Results](image-url)
Stemming

The obtained results are comparable with the ones obtained using FreeTAG approach.

Figura 5.26. Stemming Result
WordNet

In this case a considerable improvement in performances was expected. Contrarily to all expectations, the provided results are generally as good as the one supplied by the FreeTAG approach.

Figura 5.27. WordNet Results
Results Comparison

![Graph of Yes Comparison](image)

Figura 5.28. Yes Comparison
Figura 5.29. More or Less Comparison

Figura 5.30. No Comparison
5.6 shows the obtained performances.

<table>
<thead>
<tr>
<th>Label</th>
<th>FreeTAG</th>
<th>Stemming</th>
<th>WordNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>40%</td>
<td>36%</td>
<td>35%</td>
</tr>
<tr>
<td>more or less</td>
<td>18%</td>
<td>21%</td>
<td>24%</td>
</tr>
<tr>
<td>no</td>
<td>24%</td>
<td>24%</td>
<td>25%</td>
</tr>
<tr>
<td>not found</td>
<td>18%</td>
<td>19%</td>
<td>16%</td>
</tr>
</tbody>
</table>

Tabella 5.6. Average percentage of resources effectively correlated to the given one

5.6 shows how in general the performances are better than the one obtained in 5.4 (the percentage related to *yes* labels increases) even if the number of *not found* resources has increased. This seems to indicate that the previous results, with more resources, actually suffer for a missing ranking algorithm, which would bring relevant resources to the top. Without is, the advantage of retrieving many more results remains not fully exploited.

**Conclusions and Suggestions for Further Improvements**

Experimental work aimed at assessing how social tagging systems can offer a valuable alternative to typical classification and search approaches. As obtained results show, it is possible to aggregate resources by means of common tags. Thus, taken a resource, it is possible to extract a certain number of resources that are also semantically correlated to the given one, yet leaving to the users the full power (freedom) of personal classification.

Furthermore, as can be noticed looking at 5.3, the Stemming and WordNet approaches return an increasing amount of results. As checked in this work, these results are in general relevant, but they have a tendency to worsen as the search by similar resources is performed. Some possible improvements are related to a more precise evaluation of the ranking as described below.
This result are expected to give a big shake to the searching engines evolution. In fact, nowadays, searching engines are keyword based: users insert topics which they are interested in and receive as search result all the documents which contain the given keywords. On the contrary social tagging systems also facilitate the retrieve of resources by similarity by exploiting the population of custom tags inserted by users.

For example, resource #169591 which titled Imaging brain activity in conscious monkeys following oral MDMA (ecstasy), has been tagged with the keywords mdma, monkey, imaging and fmri. Within FreeTAG search results with such tags it is possible to find resources like:

- **MDMA use is associated with increased spatial BOLD fMRI visual cortex activation in human MDMA users** with associated tags mdma, imaging, human, fmri

- **Neuroimaging findings with MDMA/ecstasy: technical aspects, conceptual issues and future prospects** with associated tags sert, mdma, imaging, human

- **Plasma drug concentrations and physiological measures in ‘dance party’ participants** with associated tags neurotoxicity, mdma

- **Acute and long-term effects of a single dose of MDMA on aggression in Dark Agouti rats** with associated tags mdma, aggression, social

- **Structure of receptive fields in area 3b of primary somatosensory cortex in the alert monkey** with associated tags monkey, sensory.

These documents are all clearly related to mdma (ecstasy), while all other tags (monkey, imaging, fmri) being less frequent and popular.

To increase the amount of founded resources with some semantic correlation the Stemming and the WordNet approaches are used.
In the considered example the resources retrieved using the Stemming approach are the same obtained with FreeTAG. WordNet technique, instead, returns within the similar resources also the following ones

- *Causes and Consequences of Methamphetamine and MDMA Toxicity* with associated tags *mdma, toxicity*

- *Recent-onset ecstasy use: association with deviant behaviors and psychiatric comorbidity* with associated tags *preexisting, mdma, anxiety, depression, human*

- *Ecstasy is low in league table of major causes of deaths* with associated tags *mdma, death*

- *Death rates from ecstasy (MDMA, MDA) and polydrug use in England and Wales 1996-2002* with associated tags *mdma, death*

- *CYP2D6 deficiency, a factor in ecstasy related deaths?* with associated tags *mdma, death, enzyme*

- *Identification of the human cytochromes P450 involved in the oxidative metabolism of Ecstasy-related designer drugs* with associated tags *mdma, enzyme*

Even if the main subject is still *MDMA*, it is possible to notice the more general nature of the semantic aggregated resources.

As already mentioned, a ranking algorithm would be helpful to retrieve the actually most relevant resources first. We can refer to it as relevance of *weights*. In fact, the FreeTAG function *similarObjects* ranks each of the retrieved resources according to a floating-point *strength* of match ranging from 0 to 1.0 which depends on the number of common tags, namely the ratio between the number of the common
tags (i.e., equal tags in the given and similar resource) and the global number of tags associated to the input resource. In many analyzed results, it happen that most of the similar retrieved resources have the the same strength value (e.g. 1, indicating a perfect match with the given resource), mostly because of the small number of involved tags. This is not conceptually acceptable as it implies that all the resources having the same strength are exactly about the same subject, and, in some cases, this is not true at all.

A possible solution is to evaluate the strength in a different and more fine grained way, so allowing a more precise ranking. Unfortunately, it is difficult to find good ranking metrics which are also efficient to compute. Some examples are proposed here as investigation for future work. Now the strength is computed as the ratio between the number of the common tags (i.e. tags equal in the given and similar resources) and the global number of tags associated to the input resource. An alternative way to compute the ranking weights with the Stemming approach, could be to also take into account the ratio between the length of the given tag and the length of the stemmed one. This would probably promote resources whose stemmed tags are not changed much in comparison with the original tags, so reducing the risk of considering tags with very different meanings but with the same stem.

Another ranking approach can be used with WordNet: starting from a set of tags $T$, after having explored the set of synonyms provided by WordNet, a new set of tags $T' \geq T$ can be computed. Furthermore there will be a certain number $K$ of synonyms retrievable from $T$ and another number $J$ of synonyms retrievable from $T'$. Starting from this point, it may be possible to compute a weight factor $w_0$ as the ratio between the amount of synonyms in common, namely the intersection of the synsets, and the total amount of synonyms, namely the union of the synsets. The final step would be the computation of the strength associated to each similar resource. This could be done by computing the ratio between the intersection of $T$
and $T'$ (the number of common tags), and their union (namely the set of all possible
tags associated to the given resource), multiplied by the weight factor $w_0$. This
solution would possibly rank first those resources which have the highest number of
synonyms, but it is also probably inefficient to compute.

Alternative ways to rank resources would include the analysis of the content of
the resources themselves, though they are not always available and often difficult to
parse, the format being usually more complex than pain text.

In the final chapter the conclusions are presented, reassuming the thesis work
and the achieved results.
Conclusions

The main goal of this thesis work was to find out possible ways of aggregation of tags and resources in the specific environment represented by the data belonging to the Connotea social tagging web site.

Firstly a statistical analysis on the data has been performed to see if the considered data were in some way comparable with the one studied by [5] related to Del.icio.us. As pointed out from such analysis in the majority of the cases the two sets of data are comparable, e.g., the number of tags in each user’s tag list or the behavior of two extreme users’ tags growth, but under some aspect they are not, e.g., the growth rate of some selected tags or the addition of bookmarks to the URLs over time. This is due to the intrinsic nature of the data: Del.icio.us is a social bookmarking web service for storing, sharing, and discovering web bookmarks in general, and Connotea involves more specialized data since it provides a reference management for all researchers, clinicians and scientists. Furthermore, less users, tags and resources are available in Connotea.

After the data analysis, the work has been devoted to the study, firstly of the relationship between tags and tagged resources and successively of possible aggregations of resources, using three different approaches: FreeTAG, Stemming and
WordNet. The first part of the analysis points out how typically the returned resources have effectively been tagged with the given keyword especially considering WordNet approach. Good results have been obtained: in all cases the retrieved resources (tagged with the given keyword) were effectively related to the input one. This part of the study is useful in a conceptual aggregation view where tags related to similar resources can be aggregated to create an ontology taking the user to a new way of research no more based on keyword but on concepts.

Additionally, the results obtained from the second part of the study have evidenced a certain attitude of aggregation between resources in all the three cases of study. In particular, better results were expected from WordNet approach, and this has revealed true in 7 out of 17 considered cases. This is a good result since this method was the one that was expected to introduce a lot of noise because it considers only the synonymous of the tags without sense disambiguation. However this part of the work has highlighted how the allocation of the strength to the similar resources has in some cases to be carried out (WordNet) and in others to be refined (Stemming).

This new search technique, after the necessary improvements, could become a fundamental link in the development and diffusion of the new conception of the semantic web, born in these last few years.
Bibliografia

[1] A social analysis of tagging (or how tagging transforms the solitary browsing experience into a social one), January 2006.  
http://www.rashmisinha.com/archives/06_01/social-tagging.html

http://en.wikipedia.org/wiki/Connotea

[3] FreeTAG.  
http://en.wikipedia.org/wiki/Freetag

http://en.wikipedia.org/wiki/Folksonomy

http://www.hpl.hp.com/research/idl/papers/tags/

http://www.getluky.net/freetag/

http://www.dlib.org/dlib/april05/hammond/04hammond.html

[8] Jones, W., Phuwanartnurak, A., Gill, R., Bruce, H., Don’t Take My Folder Away! Organizing Personal Information to Get Things Done, Proceedings of


http://en.wikipedia.org/wiki/Social_bookmarking

[22] Social software.
http://en.wikipedia.org/wiki/Social_software

[23] Scuttle (about).
http://www.tamba2.org.uk/scutt/about.php

http://en.wikipedia.org/wiki/Tags


http://en.wikipedia.org/wiki/Taxonomy

[27] Top 5 social bookmarking services.
http://www.pandia.com/sew/327-social-bookmarking.html

http://www.comp.lancs.ac.uk/computing/research/stemming/general/porter.htm


http://wordnet.princeton.edu/