Extracting Knowledge from Folksonomy
フォークソノミーからの知識抽出に関する研究

by
OGINO Ken
荻野 健

A Master Thesis
修士論文

Submitted to
The Graduate School of
The University of Tokyo
on January 29, 2008
in Partial Fulfillment of the Requirements
for the Degree of Master of Science

Thesis Supervisor: SATO Hiroyuki　佐藤 周行
Associate Professor of Information Technology Center

論 文 要 旨

Web に蓄積される情報の量は，近年ますますその勢いを増してきている．それに伴い，雑多で巨大な Web という情報源から有用な情報を取り出す Web マイニングの技術はより重要になっている．そのような状況の中で，近年新たな Web マイニングの対象として Folksonomy と呼ばれる分類手法が注目され，盛んに研究が行われている．Folksonomy では，エンドユーザが主体となる点，大勢のユーザが情報を共有して共同で分類を構築していく点など，今までの情報システムとは異なる特徴を持っている．そのため，今までには得られなかった新たな情報を Folksonomy から抽出することが期待される．本研究では Folksonomy からの情報抽出として，タグの階層構造の構築とその評価，そして SVM を利用してタグ付けの機械学習を行った．階層構造の結果としては，直感的には有用な構造を得ることができた．また，評価に関しては，有意な結果を出すことは出来なかったものの，階層構造の新たな評価手法を提示することができた．機械学習の結果としては，正解率がマイクロ平均で 92%の有用な分類器を得た．また，オブジェクト当たりのタグ付け数を指標とすることで性能の良い分類器を選択することが可能になることが分かった．

目 次

図 目 次

表 目 次

第1章　序論

1.1　背景

World Wide Web(以下 Web) から有用な情報を抽出する “Web マイニング”の技術は近年ますますその重要性を増してきている．Web 上には様々な情報が日々蓄積されており，2007 年 12 月に Netcraft [1)] が公開した “December 2007 Web Server Survey”によると，現在の Web 上に存在するサイト数は一億五千万件以上にものぼる [2)]．また，一年間あたりの増加数も過去最高を記録しており，情報の蓄積になお拍車がかかっている状態である．近年ではブログの普及により個人が容易にそれぞれの持つ情報を Web 上に公開することができ，情報の多様性という面でも日々成長しているといえる．

このような巨大で雑多な情報源である Web から有用な情報を抽出する Web マイニングの研究が盛んに行われている．Web マイニングは情報検索やデータベース・自然言語処理・機械学習などの様々な分野と関連があり，非常に幅広い研究分野である．

Web マイニングは Web 内容マイニング・Web 構造マイニング・Web 利用マイニングの 3 つの分野に大きく分けられる [3)]．Web 内容マイニングは Web 文書のコンテンツに対して自然言語処理や機械学習などを応用し，有用な情報を引き出そうとする研究である．Web 内容マイニングの研究は，視点の違いからさらに 2 つの分野に分けられる．IR(Infomation Retrieval) と DB(Database) の二つである．IR は主に文書検索に主眼を置いたもので，多くの文書の中から必要なものを検索したり，フィルタリングしたりする精度の向上を目的とする研究である．DB は，Web 上のデータをモデル化して統合することで，キーワード検索に代わるより洗練されたクエリを作ることが主な目標である．Web 構造マイニングはリンクによって構成されるグラフ構造から何らかのモデルを抽出する．そのモデルに従い，Web サイトの分類やサイト同士の関連性を計算することができる．最後に Web 利用マイニングは Web そのものというよりは，Web を利用するユーザの振る舞いに着目する．サーバやクライアントに残るログなどのデータを解析し，ユーザモデリングやマーケティングを行う．

このように Web マイニングの研究が広く行われている中で，新たな Web マイニングの対象として “Folksonomy”と呼ばれるものが注目され，盛んに研究されるようになってきている．

Folksonomy とは Web 上で写真や映像・文書などの何らかのオブジェクトに対し，その内容や属性に即した「タグ」 を閲覧者が付加して分類するシステムのことである．Folksonomy の名前は,「人々」の意の “folk”と,「分類学」の意の “taxonomy”が合成されて作られたもので，その名の通り大規模な数の人間がそれぞれにタグ付けを行い，それらを集約して全体として一つの分類を作成していくのが最大の特徴である．Folksonomy は，Web の利点である大規模性や低コスト性を最大限に利用したシステムとなっており，その登場以来多くの Web サービスでの分類手法として採用されている．

そしてこの新たな分類手法が広まるにつれて，Folksonomy 自身の持つ特性を調べたり，Folksonomy を利用して新たな Web マイニングを行う研究が近年では盛んに行われている．例えば Folksonomy のタグを解析して知識構造を抽出したり，Web 検索や検索結果の個人化 (personalization) への利用，セマンティック Web と組み合わせるなど，様々な可能性が検討されている．

本研究でも Folksonomy を対象とした Web マイニングの新たな手法を提案する．提案手法は 2 つあり，まず一つ目としてタグの分布を解析してタグの意味的な階層構造を導出する手法を提案する．次に，Web ページの内容とそれに付加されたタグの関係を機械学習を用いて学習し，自動的にタグ付けを行う分類器を作成する手法を述べる．

1.2 本論文の構成

本論文の構成は以下のようになる．2 章でまず本論文の研究対象である Folksonomy について簡単に紹介する．3 章で Folksonomy からタグの階層構造を抽出する手法とその評価について述べる．4 章では機械学習の手法を用いてタグ付けの学習を行い，自動的にタグを付加できる分類器を作成してその評価を行う．最後に，5 章で本研究の結論と今後の課題を述べる．

第2章　Folksonomy

本章では研究対象である Folksonomy について述べる．まず 2.1 節では Folksonomy の持つ特徴とそれによるメリット，デメリットを説明する．次に 2.2 節では実際に Folksonomy を利用している Web サービスについて簡単に紹介する．2.3 節で一般的な Folksonomy における解析手法を説明したのち，2.4 節で Folksonomy を対象として行われている研究の紹介を行う．

2.1　Folksonomy とは

Folksonomy という名前は「人々」の意味である “folk” と「分類学」の意味である “taxonomy” を合成した言葉である [4]．その名の表すとおり，Folksonomy は多くの人々(ユーザ) によって構成される分類のこと表す．実際にはなんらかの分類対象 (以後オブジェクトと呼ぶ) に対し，そのオブジェクトの閲覧者あるいは利用者一人一人がその内容や属性に即した「タグ」を付加し，それを全体で集計，共有することで分類を行う手法のことである．また，タグをオブジェクトに付加することをタグ付けと言う．

Folksonomy には次のような点で従来の分類法とは異なる．

- エンドユーザが主体となって分類を行う
- 分類に参加する人の数が大規模
- 分類の際に全体の構造を考えない

従来の分類の多くは一部の権威者や専門家 (Authority) によって作成されたものである．例えば Yahoo [5] が以前まで使用していたディレクトリ型の検索エンジンや，生物の樹形図などは Authority による分類であると言える．しかし Folksonomy の場合は Authority ではなくエンドユーザが主体となってタグ付けを行うのが特徴的である．

また，従来では分類を行うのは少数の人間か，人数が多い場合でも一つの組織によって行われる事が多い．そのため，ある一つの価値観に基づいて分類が行われる．しかし Folksonomy では大多数のユーザを集めてタグ付けが行われるため，多種多様な価値観が入り混じることになる．

最後に，従来の分類は全体の構成が考えられて作られている．しかし Folksonomy の場合は各ユーザがそれぞれの好きなようにタグ付けしているだけである．

2.1.1　メリット

Folksonomy のメリットとしては以下が挙げられる．

- 分類にかかる手間が少ない
  Folksonomy ではユーザにかかる手間は非常に小さい．タグは任意の文字列でよいため思いついたままに付加することができる．また，各ユーザはそれぞれ自分にとって都合のよい分類 を作ればよく，全体の分類について考慮する必要がないため難しいことを考える必要はない．

- 分類に多様性がある
  一つの価値観で分類を行う従来の分類に比べて，Folksonomy では多くのユーザがタグ付けを行うため，様々な観点からの分類を得ることができる．これにより今までには気が付かなかった意外な関係性や概念を発見できる可能性がある．

- 柔軟性が高い
  Folksonomy を全体としてのオブジェクトの分類は，そのオブジェクトに付加されたタグのセットによって多数決的に決まる．すなわちタグの付き具合によって動的に分類が変化するため，柔軟性に優れている．

- 集約した情報が得られる
  例えば集計期間内で多くタグ付けされたオブジェクトを，現在注目のオブジェクトとしてユーザに紹介するなど，新たな情報を得ることができる．

2.1.2 デメリット

Folksonomy のデメリットは次のような点が挙げられる．

- 分類の信頼性が低い

- 曖昧な部分が存在する

ユーザの間違った知識に従って分類されてしまうことが起こるため，分類の正しさは保証できない．多くのユーザのタグ付けを集計することで，一人のユーザの間違いの効果を薄くすることができるが，専門的な知識を要する分類を正しく行うことは難しい．また，故意に事実とは異なるタグ付けを行ってユーザを誘導するようなスパム行為が行われる可能性もある．現状ではまだスパムは大きな問題となっていないが，いずれ解決するべき問題であると言える．

次に，多数のユーザがそれぞれの語彙を使ってタグ付けを行っていくため，曖昧性が生じてしまうという問題もある．同じタグだが使うユーザによって意味が異なる曖昧性 (ambiguity) や，異なるタグだが意味は同じであるシノニム (synonym) などが挙げられる．このような曖昧性に関する問題を解決するための研究も行われている [6]．

2.2 Folksonomy を利用したサービス

本節では実際に Folksonomy を利用しているサービスを紹介する．Folksonomy は主に管理対象としているオブジェクトで分類することができる．**表 2.1** に主な管理対象と対応するサービスの例を示す．本節では表 2.1 の中でも特にメジャーな del.icio.us [7] に主眼を置いて紹介する．

del.icio.us はブックマーク共有サービスの一つである．ブックマーク共有サービスとは Web 上にブックマークを保存・管理できるサービスで，ユーザは自分のブックマークにタグをつけて管理する．各ユーザのタグ付け情報は全体で共有され，検索などに用いることができる．**図 2.1** に del.icio.us のトップページを示

表 2.1: Folksonomy を利用したサービスの例

対象	サービス
Web 文書	del.icio.us，digg [8], はてなブックマーク [9]
画像，動画	Flickr [10]，YouTube [11]
論文	CiteULike [12]，Connotea [13]

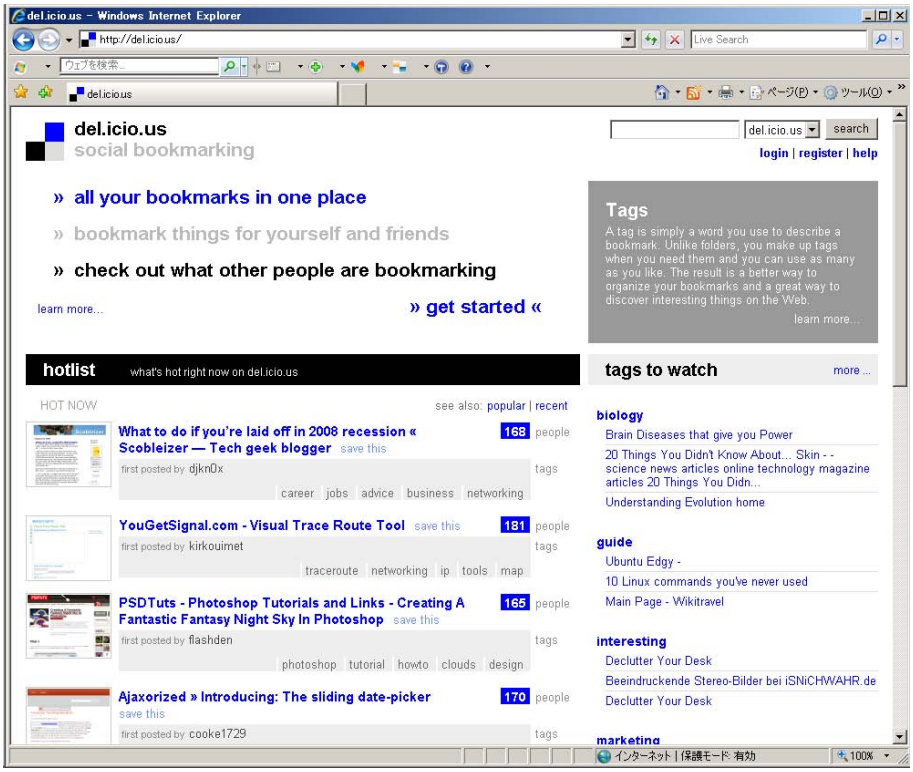


図 2.1: del.icio.us のトップページ

す．トップページではユーザのタグ付け情報を集計し，現在人気の Web ページをランキング形式で表示している．

図 2.2 は，del.icio.us で管理されているブックマークの一つを詳細ページで見たものである．右上の方にこのブックマークに付加されているタグの一覧が表示されている．また，その下にはブックマークを行ったユーザと使用したタグの一覧が表示されている．これらの情報は公開されており，ブックマークに使われているタグの一覧を眺めてブックマーク先の Web ページの内容が自分の興味に合うかどうか推測することなどができる．

図 2.3 は del.icio.us でタグ付けを行うときの画面である．図内の “tags” の欄に自分なりのタグを入力して登録を行う．下の方にある “popular tags” は，このブックマークにすでに付加されているタグで使用頻度が高いものを候補として例示している．ユーザは自分で考えたタグだけでなく，例示されたタグの中から選んでタグ付けすることも多い．

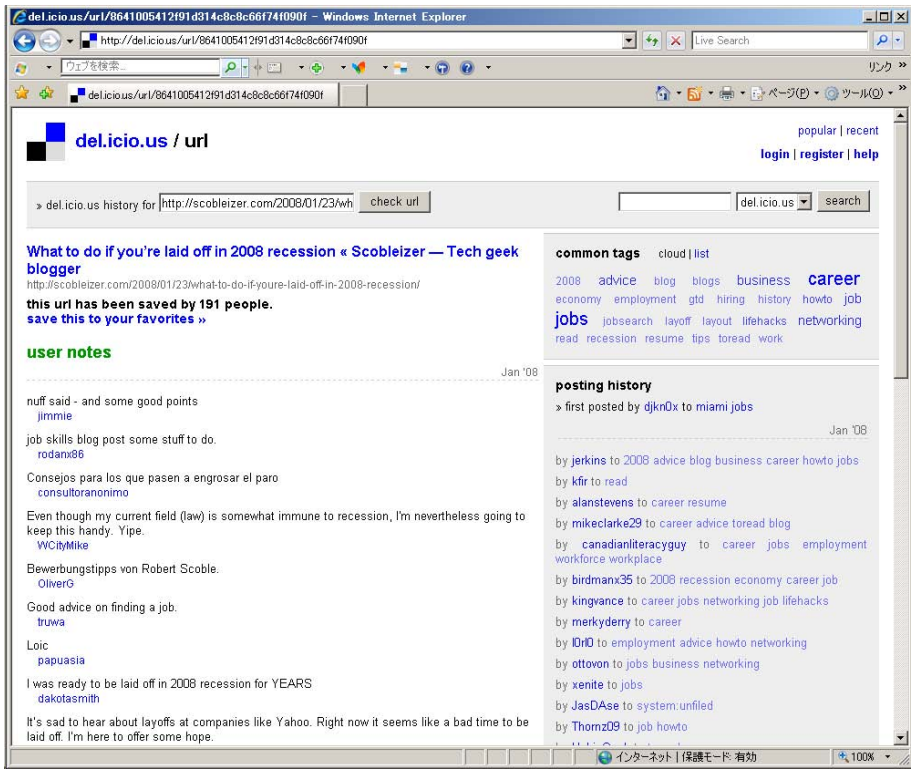


図 2.2: del.icio.us のブックマークの詳細ページ

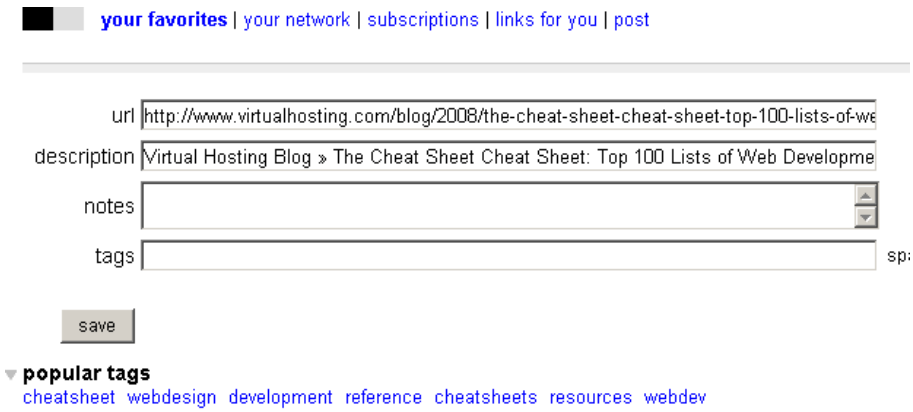


図 2.3: del.icio.us のタグ付けを行うページ

2.3 一般的な解析手法

現在ある Folksonomy を利用したサービスで，タグ付けを構成する要素は 4 つある．すなわちタグ付けが行われた時間とそれを行ったユーザ，対象となるオブジェクト，用いられたタグの 4 つである．従って，一つのタグ付けは次の四つ組で表すことができる．

$$(user, object, tag, time)$$

しかし，多くの既存研究ではタグ付けを行った時間は考慮していない．本研究でも時間情報に関しては特に利用していないので，今後は特に言及しないこととする．なお，時間情報を利用して Folksonomy の時系列的な分析を行っている研究はいくつか存在する [14, 15]．

さて，時間情報を無視するとタグ付けは次の三つ組で表現される．

$$(user, object, tag)$$

ユーザの集合を $U = \{u_1, u_2, ..., u_I\}$，オブジェクトの集合を $O = \{o_1, o_2, ..., o_J\}$，タグの集合を $T = \{t_1, t_2, ..., t_K\}$ とする．そして次のような 3 階のテンソル $A \in R^{I \times J \times K}$ を定義する．

$$\begin{aligned} A_{ijk} &= 1 \quad if \quad (u_i, o_j, t_k), \\ &= 0 \quad otherwise \end{aligned}$$

すなわち，あるユーザ u_i がオブジェクト r_j に対しタグ t_k を付加するタグ付けが行われた場合 $A_{ijk} = 1$ となる．このようにすることでタグ付けの情報が全てこのテンソル A で表現できる．

Folksonomy の解析を行う場合，共起数に注目するのが常套手段である．例えば，ある二つのタグが共に一つのオブジェクトに対してタグ付けされていた場合，それらのタグはオブジェクトを介して共起していると言う．オブジェクトを介したタグ同士の共起は次の行列 $B \in R^{J \times K}$ で表現できる．

$$B = \bigvee_i A_{ijk}$$

ここで，$\bigvee_i$ はテンソル A を i の次元に関して論理和をとるものであるとする．行列 B はオブジェクトと，そのオブジェクトにタグ付けされたタグの対応関係を表す行列に相当する．したがってあるオブジェクト r_j にタグ t_k が付加されていた場合，$B_{jk} = 1$ となる．

よく共起するタグ同士はよく似た意味を持つと推測できる．そこで，共起の回数をタグ間の類似度の指標に用いることが多い．あるタグ t_{k_1} と t_{k_2} の間の共起数は次の式から計算できる．

$$c_{k_1,k_2} = b_{k_1}^{tr} \cdot b_{k_2}$$

ここで b_k は行列 B の k 番目の列ベクトルである．

以上のように，基本的な Folksonomy の研究ではユーザ，オブジェクト，タグの共起に着目して行われることが多い．

2.4 研究紹介

2.4.1 検索への応用

Folksonomy の情報を利用して Web ページの検索を行う研究が行われている．これらの多くは PageRank [16] アルゴリズムをベースにしている．Bao らは PageRank を Folksonomy 用に修正した SocialPageRank を提案している [17]．SocialPageRank のアルゴリズムを**図 2.4** に示す．ここで，M_{UO} はユーザがどのオブジェクトにタグ付けしたかを保持する行列である．同様にして M_{OT} はオブジェクトにどのタグが付いている

Algorithm: SocialPageRank

Step1 Input:

$$M_{UO} = \bigvee_k A_{ijk}, \quad M_{UO} \in R^{I \times J}$$
$$M_{OT} = \bigvee_i A_{ijk}, \quad M_{OT} \in R^{J \times K}$$
$$M_{UT} = \bigvee_j A_{ijk}, \quad M_{UT} \in R^{I \times K}$$

random initial SocialPageRank score $O_0 \in R^J$

Step2 Do:

$$U_i = M_{UO} \cdot O_i$$
$$T_i = M_{UT}^{tr} \cdot U_i$$
$$O'_i = M_{OT} \cdot T_i$$
$$T'_i = M_{OT}^{tr} \cdot O'_i$$
$$U'_i = M_{UT} \cdot T'_i$$
$$O_{i+1} = M_{UO}^{tr} \cdot U'_i$$

Until O_i converges

Step3 Output:

P^*: the converged SocialPageRankScore

図 2.4: Social Page Rank のアルゴリズム

かを保持する行列，M_{UT} はユーザがどのタグを使っているかを保持する．また，O_0 が各オブジェクトの SocialPageRank の値を保持する初期ベクトルである．

Folksonomy を利用した検索には二つ特長がある．一つ目は Web のエンドユーザからの視点でランク付けされるという点である．これは，今までの PageRank が Web ページの管理者の視点からランク付けしているのと比べて特徴的である．もう一つの特徴は，必ずしもキーワードが必要ないという点である．これは，2.3 節で述べたようにタグの類似度を共起数から計算することができるため，検索キーワードそのものがマッチする必要がないためである．今までの検索アルゴリズムがキーワードが一致しないと結果に現れないのと比べて大きな利点であるといえる．

2.4.2 セマンティック Web への応用

セマンティック Web とは Web 上の情報の「意味」を人間だけでなく，マシンにも処理できるようにする次世代の Web である [18]．例えば人間の場合，Web ページを見ればページ中のある数字が値段を表しているのか個数を表しているのかの意味が理解できるが，マシンが HTML を解析しても与えられた数字の意味を理解することは難しい．そこであらかじめオントロジーで知識ベース (knowledge base) を構成しておき，その情報を Web ページにメタデータとして埋め込むことでマシンでも情報の意味を理解して処理することが可能になる．

このメタデータを埋め込むことをアノテーションと言う．しかしアノテーションを行うにはそれなりの専門知識を必要とし，さらに Web ページの管理者自身がアノテーションを行うのは面倒なため，コストが大きい．それがセマンティック Web がまだ Web 上で普及しない大きな要因の一つとなっている．

このような背景のもとに，Folksonomy を利用してアノテーションのコストを小さくしようとする研究が行われている．Folksonomy のタグ付けは専門的な知識は必要ないためコストが小さい．そこでとりあえず最初は Folksonomy のタグ付けを行っておき，後でタグとオントロジーの概念を結びつける手法が提案されている [19, 20]．

また他にもセマンティック Web と Folksonomy を組み合わせた研究として，オントロジーをタグの情報を使って修正していくことで柔軟性を持たせようという研究もなどある [21]．

第3章　タグの階層構造抽出

3.1　背景

通常の Web のキーワード検索には「ナビゲーション」機能を有するものもある．ナビゲーションは，例えば検索したいものがあるのに検索キーワードがはっきりと思い出せない時や特に主たる目的もなく漠然と自分の興味を引くものを探しにたい時など，正しい検索クエリがはっきりと分からない時に使われ，検索キーワードに関連するようなキーワードをユーザに提示して徐々に検索対象を絞り込んでいく機能である．

Folksonomy におけるタグも，あらかじめタグの意味に応じて階層構造を作成しておくことで，検索時にユーザをナビゲーションすることが期待できる．

Folksonomy からタグの階層構造を抽出する研究はいくつか存在する．Heymann らはタグ間の類似度を計算し，類似度の高いペアから順に階層関係をつくることでナビゲーションに使えるようなタグの階層構造を構築できると主張している [22]． Begelman らはタグ同士の関係をタグ間の共起数で辺に重み付けしたグラフ構造で表した [23]．このタグの類似度グラフをグラフ分割の手法を利用して関係性の低い辺から切り離していき，最終的に類似タグのクラスタを作る手法を提案している．また，Grahl らは k-means 法を利用して階層構造を構築した [24]．この手法では，まずタグ全体の集合を k-means 法でクラスタに分割した後，各クラスタからタグを一つ選んで上階層のタグとする．次に，選ばれた上階層のタグに対してもう一度 k-means 法でクラスタリングを行い，各クラスタを代表するタグをまた一つ選んで最上段の階層とする．最大 3 段からなる階層構造を Grahl らは提案している．

上記の既存研究では全てなんらかのクラスタリング・アルゴリズムを利用してタグの階層構造を作成している．本研究でもクラスタリング利用して階層構造の抽出を行う．本研究の特徴として，タグの「サイズ」を考慮して階層構造を作成した．

また多くの既存研究において，作成したタグの階層構造の評価は行われていない．これは階層構造の評価には人間の「知性」が関連してくるために評価が難しく，まだ評価方法が確立されていないことによる．そこで本研究では，Open Directory Project (ODP) [25] を利用して，階層構造の評価を試みた．

3.2　関連研究

3.2.1　畳み込みカーネル

本節では階層構造の評価に用いる，畳み込みカーネルによる木の共通部分構造の数え上げについて説明する．

畳み込みカーネルは二つの構造データ T_1 と T_2 が与えられたとき，

$$K(T_1, T_2) = \sum_{s_1 \in S(T_1)} \sum_{s_2 \in S(T_2)} K^S(s_1, s_2) \tag{3.1}$$

と定義される [26]．ここで，$S(T)$ は T から取り出せる部分構造の集合を表し，K^S は二つの部分構造の間に定義されるカーネル関数であるとする．式 (3.1) の意味するところは，構造データ全体のカーネル関数はその部分構造間のカーネル関数を全て足し合わせることで得られるということである．K^S がカーネル関数であるならば，式 (3.1) もカーネル関数であることが保証される．

T_1 と T_2 が共に木構造であった場合，式 (3.1) は以下のように変形できる．

$$K(T_1,T_2) = \sum_{v_1\in V(T_1)}\sum_{v_2\in V(T_2)}\sum_{s_1\in S_{v_1}(T_1)}\sum_{s_2\in S_{v_2}(T_2)} K^S(s_1,s_2) \tag{3.2}$$

$$= \sum_{v_1\in V(T_1)}\sum_{v_2\in V(T_2)} K^R(v_1,v_2) \tag{3.3}$$

ここで，$V(T)$ は木 T における頂点の集合，$S_v(T)$ は頂点 v を根として持つような T の部分木の集合とする．また $K^R(v_1,v_2)$ は $S_{v_1}(T_1)$ と $S_{v_2}(T_2)$ に限定したときのカーネルで

$$K^R(v_1,v_2) = \sum_{s_1\in S_{v_1}(T_1)}\sum_{s_2\in S_{v_2}(T_2)} K^S(s_1,s_2)$$

とする．このように木構造に関するカーネル関数を木カーネルと言う．

鹿島らは T_1 と T_2 が順序付きラベル木である場合に，式 (3.3) の $K^R(v_1,v_2)$ を以下のように定義した木カーネルを設計した [27]．

- v_1 あるいは v_2 が葉のとき

$$K^R(v_1,v_2) = L(v_1,v_2) \tag{3.4}$$

- v_1 と v_2 が葉でないとき

$$K^R(v_1,v_2) = L(v_1,v_2)\cdot \bar{K}^R_{v_1,v_2}(\#ch(v_1),\#ch(v_2)) \tag{3.5}$$

$$\begin{aligned}\bar{K}^R_{v_1,v_2}(i,j) = {} & \bar{K}^R_{v_1,v_2}(i-1,j) + \bar{K}^R_{v_1,v_2}(i,j-1) \\ & -\bar{K}^R_{v_1,v_2}(i-1,j-1) \\ & +\bar{K}^R_{v_1,v_2}(i-1,j-1)\cdot K^R(ch(v_1,i),ch(v_2,j))\end{aligned} \tag{3.6}$$

ここで，$L(v_1,v_2)$ は頂点 v_1 と頂点 v_2 についているラベルが等しければ 1 を，等しくなければ 0 を返す関数，$\#ch(v)$ は頂点 v の持つ子頂点の数を表す．また，境界条件 $\bar{K}^R_{v_1,v_2}(i,0) = \bar{K}^R_{v_1,v_2}(0,j) = 1$ とする．

上記の再帰式で定義された木カーネルは，二つの木間の共通部分木構造の数をかぞえあげるものに等しいことが分かっている [28]．再帰式を動的計画法を用いて計算することで，木カーネル $K(T_1,T_2)$ を $O(|V_1||V_2|)$ の計算量で計算することができる．

3.3 提案手法

3.3.1 理想的なモデル

ここではタグの階層構造抽出のモデルとなる概念を述べる．**図 3.1** は Folksonomy でのタグ付けの概略図である．各オブジェクトに対し，ユーザ A はそれぞれに異なるタグ (“Win”，“Mac”，“Linux”) をタグ付けしている．対してユーザ B は全てのオブジェクトに対して同じ “OS” というタグをタグ付けしている．このように，ユーザごとに持っている興味や知識の差によって，タグ付けの挙動も変化すると考えられる．図のユーザ A には各オブジェクトごとの細かな違いが重要となるが，ユーザ B はより抽象的な側面に注目しているため異なる抽象度のタグがタグ付けされることになる．多くのユーザが同じオブジェクトに対してタグ付けした場合，そこには様々な抽象度のタグが存在すると考えられる．従って，同じオブジェクトに関して共起している 2 つのタグ間には階層関係が存在している可能性がある．特に，頻繁に共起するタグのペアは階層関係を持つ可能性が高いといえる．

タグの共起率によって階層構造の候補となるタグのペアが与えられたら，次にそのタグ間の上下関係を判定する必要がある．この上下関係は，タグの “サイズ” を比較することで判定する．ここでタグのサイズとは，

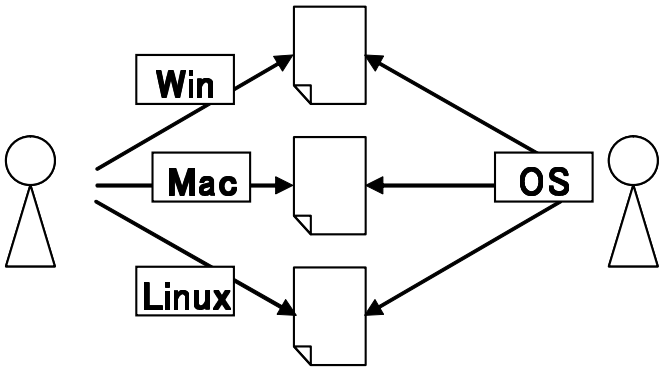


図 3.1: タグ付けの概要

そのタグがタグ付けされているオブジェクトの数のことを言う．本研究では，より抽象的な意味を持つタグのほうがそのサイズが大きくなると仮定した．例えば図の場合，タグ “Win”や “Mac”，“Linux”などはそれぞれ一つのオブジェクトにのみタグ付けされているのでそのサイズは 1 である．そしてタグ “OS”は 3 つのオブジェクトに対してタグ付けされているためそのサイズは 3 である．ここで，タグ “OS”は他のタグよりもより抽象的な意味を持っているためより多くのオブジェクトに対してタグ付けされる機会が多く，サイズも大きくなると考えられる．従って，共起率の高い二つのタグが与えられた時に，よりサイズの大きいほうのタグが親タグであると推定することができる．

親タグはその子となるタグ同士の関連性に大きな影響を与える．例えば親タグが “computer”であった場合，“Mac”と “OS”は高い関連性 (高い共起率) を持つだろう．一方で親タグが “Food shop”であった場合，“Mac”と “OS”間の関連性は下がり，逆に “Mac”と “Fast-food”というタグの関連性は上がると考えられる．このような親タグの影響はタグにおける文脈の変化であると捉えられる．タグ間の関連性を考えるときは，このタグの文脈を考慮する必要がある．

3.3.2 手法

ここでは 3.3.1 節で述べたモデルに基づいて，タグの階層構造を抽出する実験を行った．ここではその手順について説明する．本研究ではまず最初にタグの共起率に基づいて関連性の高いタグを集めたクラスタを作成する．そして，各クラスタ内でタグの階層構造を構築するという手順になる．

クラスタの作成

共起率に基づいてタグ間の関連性を評価し，関連性の高いタグを集めたクラスタを作成する．評価の指標として，AEMI(Augmented Expected Mutual Information[29]) を用いた．AEMI は次式で表される．

$$\begin{aligned} \mathrm{AEMI}(a,b) &= \mathrm{MI}(a,b) + \mathrm{MI}(\overline{a},\overline{b}) \\ &\quad -\mathrm{MI}(a,\overline{b}) - \mathrm{MI}(\overline{a},b) \qquad (3.7) \\ \mathrm{MI}(a,b) &= P(a,b)\log\frac{P(a,b)}{P(a)P(b)} \qquad (3.8) \end{aligned}$$

ここで，$P(a)$ はタグ a がオブジェクトに対してタグ付けされる確率を表す．そして，$P(a,b)$ はオブジェクトに対してタグ a と b の両方がタグ付けされる確率である．さらに，$P(\overline{a})$ はタグ a がオブジェクトに対し

てタグ付けされない確率を表す．MI(Mutual Information [30, 31]) は共起率を評価するための一つの手法であり，AEMI はこれを基にしている．AEMI の値は二つのタグ a と b の両方があるオブジェクトに対してタグ付けされているか，両方ともタグ付けされていなければ高くなる．逆に a と b のどちらか片方のみがあるオブジェクトにタグ付けされていると低くなる．本研究ではこの AEMI を用いて次のようにクラスタ C を作成した．

$$\begin{aligned} tags &= \{t_0, t_1, t_2, ..., t_n\}, \\ \forall t_r \in tags, C_{t_r} &= \{t | t \in tags, \\ &\quad \mathrm{AEMI}(t_r, t) > threshold, \\ &\quad size(t_r) > size(t)\}, \end{aligned}$$

ここで，$size(t)$ はタグ t のサイズ (タグ t によってタグ付けされているオブジェクトの数) を表す．クラスタの作成にはまず最初にルートタグ t_r を選ぶ．このルートタグ t_r は，クラスタ内で作られる階層構造のルートとなるタグである．クラスタ C_{t_r} にはこの t_r との AEMI の値が閾値より大きく，サイズが t_r より小さいタグが含まれる．このクラスタリングにおいて，異なるクラスタ内ならば同じタグが重複して存在することを許した．

階層構造の構築

作成したそれぞれのクラスタ C_{t_r} においてタグの階層構造を作成した．**図 3.2** に階層構造作成のアルゴリズムを示す．ここでは，先ほど述べた AEMI に少し修正を加えた指標を用いている．新しい指標は以下の式で表される．

$$\begin{aligned} \mathrm{AEMI}(a, b|t_r) = \mathrm{MI}(a, b|t_r) + \mathrm{MI}(\overline{a}, \overline{b}|t_r) \\ -\mathrm{MI}(a, \overline{b}|t_r) - \mathrm{MI}(\overline{a}, b|t_r), \end{aligned} \tag{3.9}$$

$$\mathrm{MI}(a, b|t_r) = P(a, b|t_r) \log \frac{P(a, b|t_r)}{P(a|t_r)P(b|t_r)}, \tag{3.10}$$

$$P(a|t_r) = \frac{P(a, t_r)}{P(t_r)}, \tag{3.11}$$

ここで，$P(a|t_r)$ はタグ t_r にタグ付けされたオブジェクトにタグ a がタグ付けされている確率を表す．図 3.2 のアルゴリズムでは，ルートタグ t_r のもとでのタグ間の共起率を評価したいので，この修正した AEMI を用いている．

図 3.2 では，まずクラスタのタグ集合内で全てのタグ間の関連性を式 (3.9) で評価する．その中で最も値の大きなタグのペア t_i と t_j を選び出し，結合して新しいタグ t_{new} をつくる．結合は次の操作で行われる．t_i と t_j のそれぞれのサイズを比較し，より大きいほうを親とする階層構造を構築してこれを t_{new} とする．クラスタのタグ集合から t_i と t_j は取り除かれ，代わりに t_{new} を加える．この操作を，選び出された t_i と t_j の式 (3.9) での評価値がある閾値を下回るまで行う．ここでは閾値にクラスタ作成のときと同じ値を用いた．

3.4 実験

3.4.1 データ

本実験で使用したアノテーションのサンプルデータは del.icio.us 上を 2007 年 2 月から 2007 年 12 月までの期間にクローリングして得たものを用いた．サンプル数は**表 3.1** に示す通りである．本実験ではこの中から使用頻度の高い上位 100 個のタグに対して，上記の手法を用いて階層構造を構築した．使用頻度上位 100 のタグを**表 3.2** に示す．

foreach C_{t_r}
  $tags = \{t | t \in C_{t_r}\}$
  while true
    find $t_i, t_j = \{\text{AEMI}(t_i, t_j | t_r) > \text{AEMI}(t_x, t_y | t_r) \quad \textbf{for all } t_x, t_y \in tags\}$
    if $\text{AEMI}(t_i, t_j | t_r) < threshold$
      break
    end if
    $t_{new} = combine(t_i, t_j)$
    $tags = tags - \{t_i, t_j\}$
    $tags = tags + t_{new}$
  end while
end foreach

図 3.2: タグの階層構造構築のアルゴリズム

表 3.1: サンプルデータの数

項目	数
ユーザ数	947,605
オブジェクト数	433,540
タグ数	1,045,630
タグ付け数	102,153,502

3.4.2 パラメータ

閾値として全 AEMI 値の平均と中央値の二つを採用した．それぞれの値は平均値が 0.00805, 中央値が 0.00530 である．

3.4.3 実験結果

実験結果は各タグのツリーで得られた．**図 3.3** と**図 3.4** にその結果を示す．図 3.3 は得られたツリーの深さの分布を，図 3.4 はツリーの要素数の分布を表したものである．それぞれの図は閾値に平均値を採用した結果 (avg) と中央値を採用した結果 (med) を並べて表示している．

図 3.3 より，ツリーの最大の深さは avg が 8，med が 10 である．. また，平均を取ると avg が 4.59，med が 5.16 と med のほうが全体的にツリーが深い．med のほうが閾値が小さく，それだけクラスタに含まれるタグの数も増えるためである．図 3.4 では，avg と med の両方とも全体の約半分程度が要素数 5 以下のツリーで構成されていることがわかる．ここも同様に平均値を取ると avg が 18，med が 25.7 である．

また，avg と med の両者共に全体の約 $\frac{1}{3}$ のツリーの深さが 1 となっている．これらは今回設定した閾値が高すぎたためにクラスタを形成できなかったタグである．

実験の結果得られたツリーは本論文の最後に添付する (Appendix)．またいくつかのツリーの例を**図 3.5**, **図 3.6**, **図 3.7**, **図 3.8**, **図 3.9**, **図 3.10** に示す．それぞれの図の左が avg，右が med のツリーである．図 3.5 ~ 図 3.9 は全体の中でも直感的に「良い」構造をしていると思われるツリーである．これらの例では，我々の立てたモデルがうまく適合しているのではないかと考えられる．

閾値が一つに固定されているために，直感的に正しくない結果になってしまっているものも見受けられる．例えば図 3.10 は，“windows” タグに “linux” や “mac” といったタグが子として存在している．これらは直感

表 3.2: 使用頻度上位 100 タグ

1	design	26	opensource	51	tutorials	76	travel
2	software	27	mac	52	html	77	rss
3	tools	28	windows	53	shopping	78	diy
4	reference	29	fun	54	education	79	tool
5	web	30	search	55	lifehacks	80	research
6	programming	31	flash	56	ruby	81	article
7	webdesign	32	internet	57	books	82	interesting
8	css	33	cool	58	photoshop	83	imported
9	blog	34	blogs	59	teck	84	mp3
10	web2.0	35	business	60	photo	85	tv
11	howto	36	technology	61	apple	86	maps
12	tutorial	37	games	62	social	87	gtd
13	free	38	humor	63	photos	88	library
14	system:unfile	39	security	64	online	89	wiki
15	linux	40	productivity	65	community	90	writing
16	ajax	41	freeware	66	media	91	code
17	javascript	42	inspiration	67	politics	92	history
18	video	43	osx	68	toread	93	hardware
19	music	44	java	69	computer	94	ubuntu
20	art	45	webdev	70	culture	95	hacks
21	tips	46	graphics	71	resources	96	wordpress
22	development	47	firefox	72	audio	97	utilities
23	photography	48	funny	73	download	98	language
24	news	49	science	74	images	99	python
25	google	50	php	75	rails	100	health

的には並列に並ぶべきである．今回の実験で採用した閾値が “windows”タグのクラスタ内では低かったために，このような結果になってしまったと考えられる．逆に閾値が高すぎるのも良くない．図 3.8 はそれぞれ avg と med での，“opensource”をルートとするツリーであるが，閾値の低い med でのツリーのほうが avg のツリーと比べてより充実したものになっている．このように，個々のクラスタ内で適切な閾値は異なると考えれるため，閾値を固定値ではなくなんらかの計算式からクラスタごとに計算することでよりツリーを改善できる可能性がある．

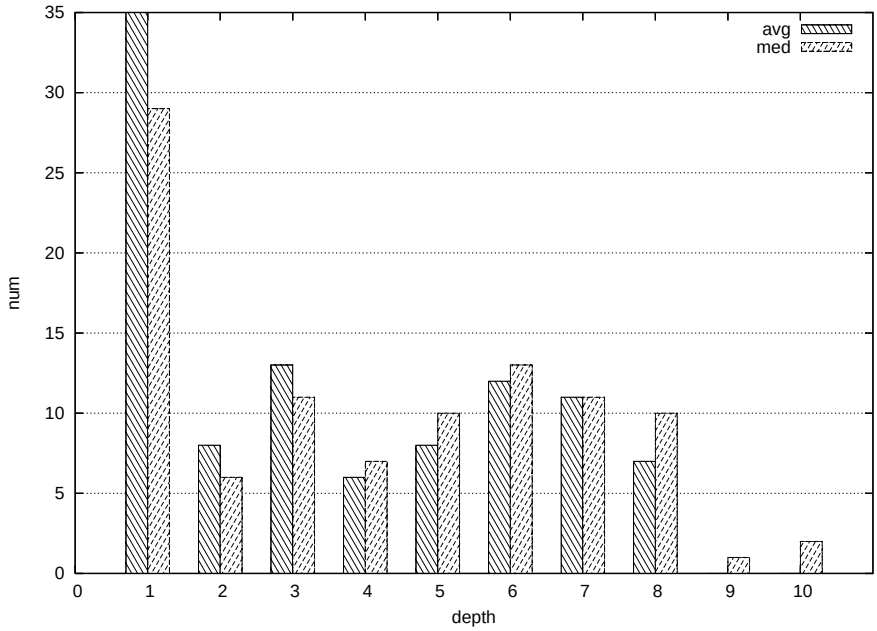


図 3.3: タグ・ツリーの深さの分布

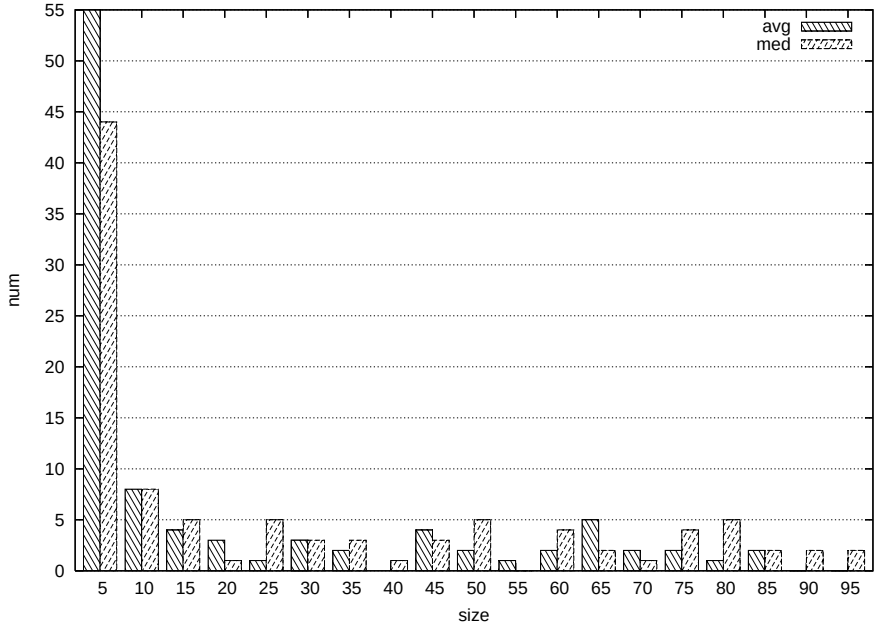


図 3.4: タグ・ツリーの要素数の分布

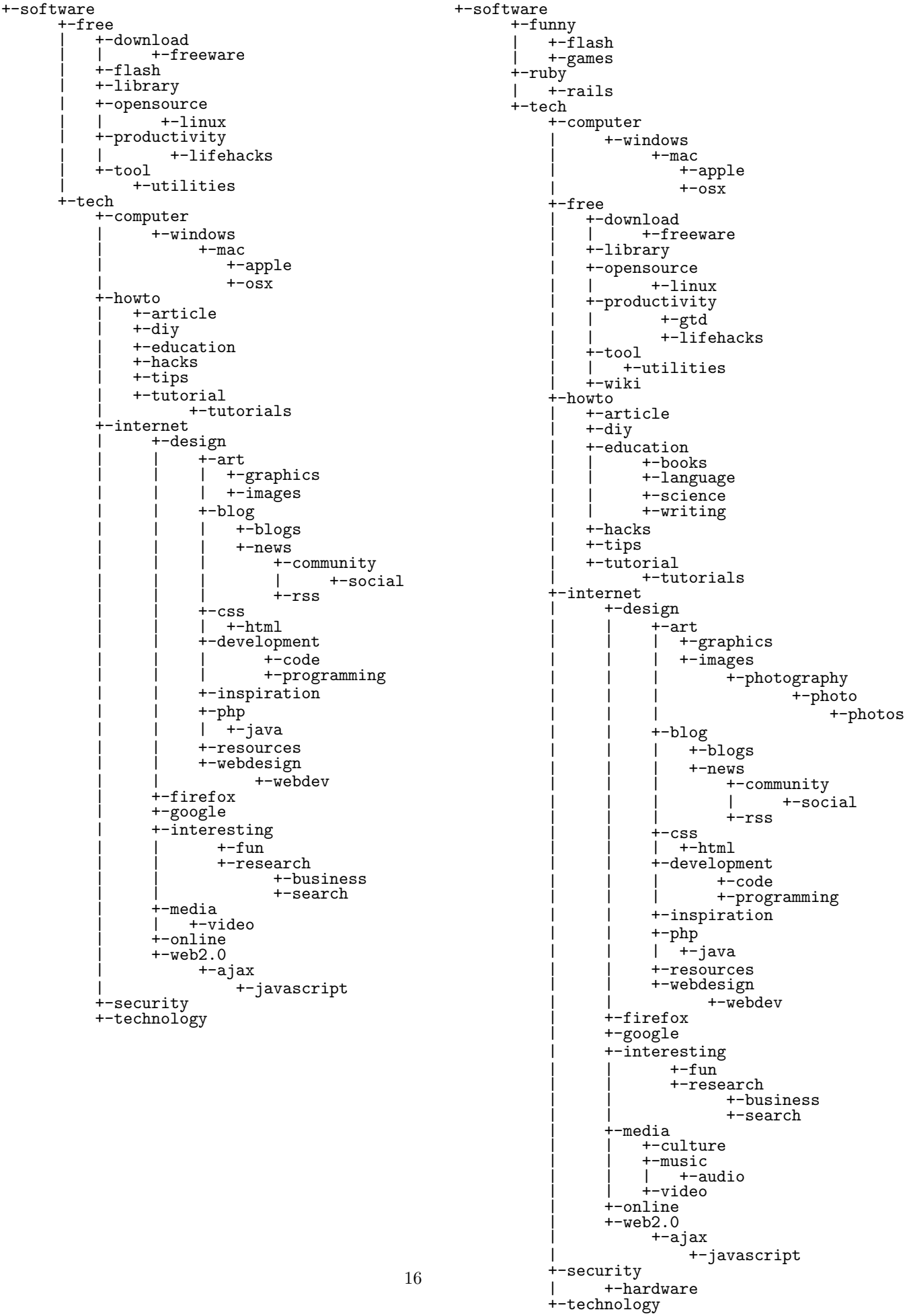


図 3.5: avg(左) と med(右) の “software”のタグ・ツリー

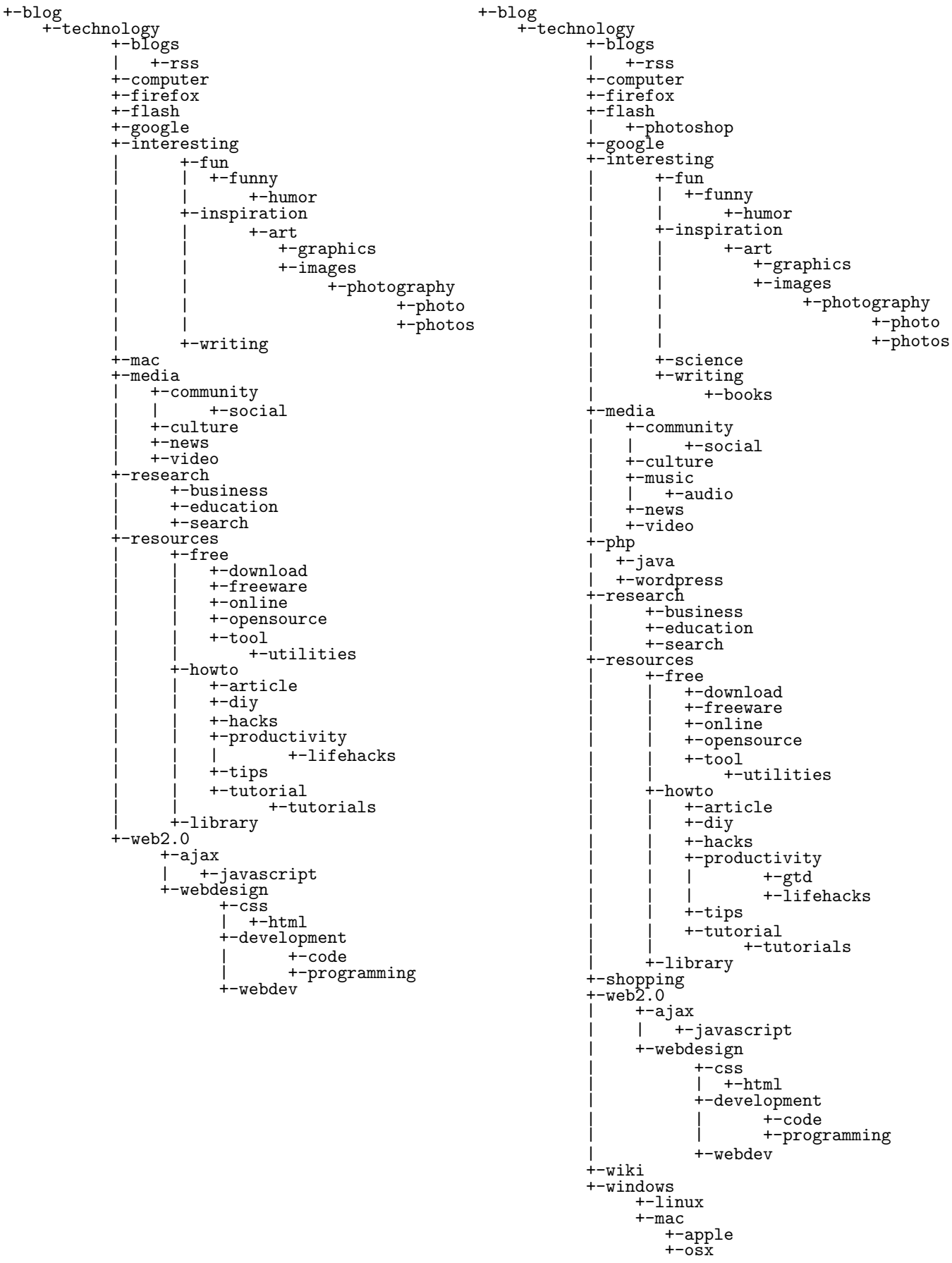


図 3.6: avg(左) と med(右) の “blog”のタグ・ツリー

図 3.7: avg(左) と med(右) の “utilities”のタグ・ツリー

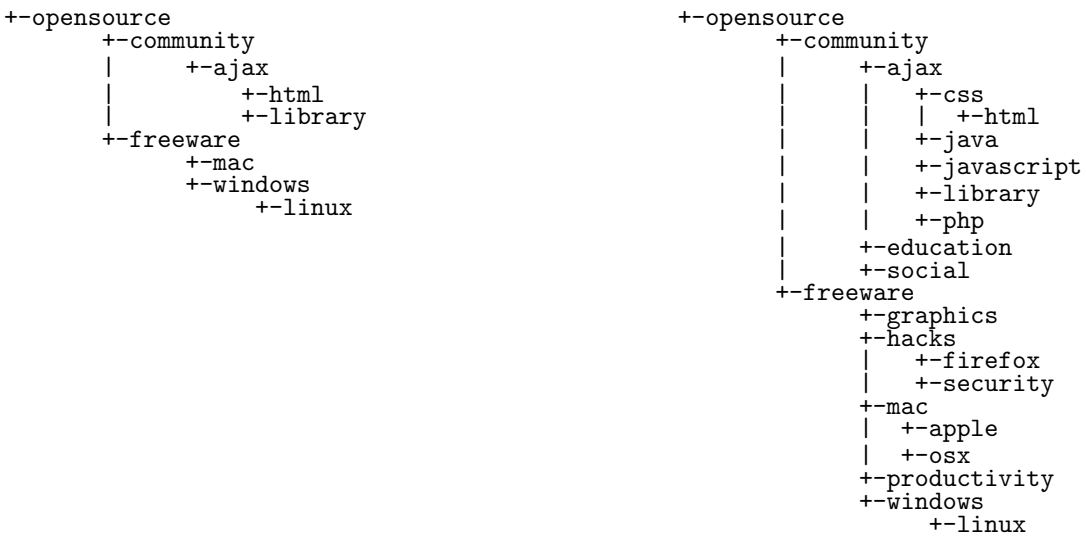


図 3.8: avg(左) と med(右) の “opensource”のタグ・ツリー

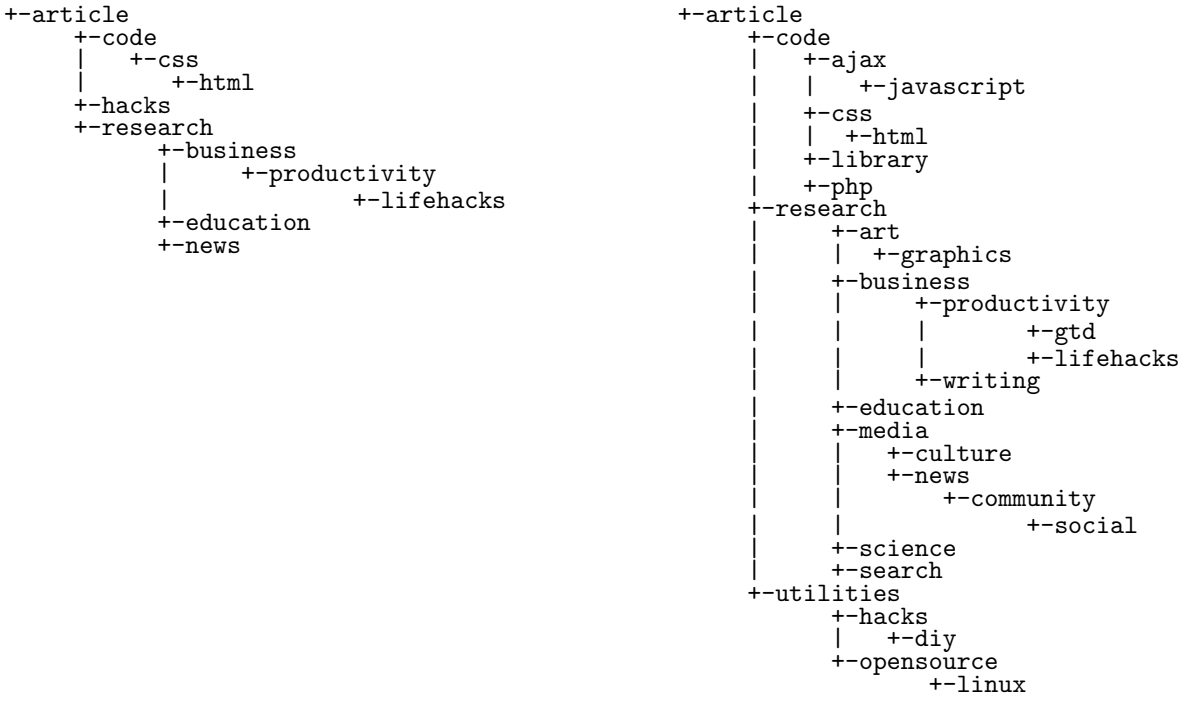


図 3.9: avg(左) と med(右) の “article”のタグ・ツリー

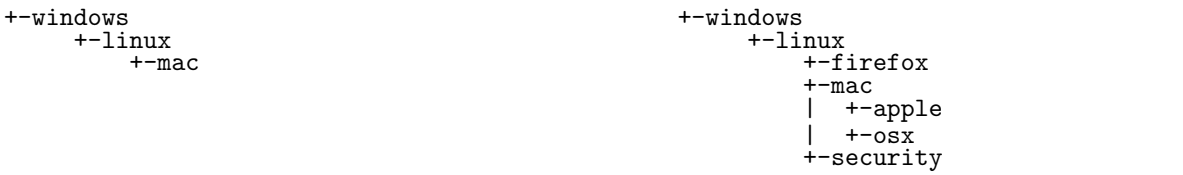


図 3.10: avg(左) と med(右) の “windows”のタグ・ツリー

3.5 評価

ここでは実験で得たタグの階層構造を評価する．本研究では，Open Directory Project における階層構造と本研究で得た階層構造を比較して評価を行う．

3.5.1 Open Directory Project

Open Directory Project はボランティア方式で運営される世界最大の Web ディレクトリである．ODP では Web サイトがディレクトリ状に構成されたカテゴリに分類される．各カテゴリごとに担当のエディタと呼ばれる管理人が割り当てられ，カテゴリにふさわしい Web サイトを登録したり，サブカテゴリを作って分類することを業務とする．**図 3.11** は，ODP のディレクトリ構造の一部を示している．

ODP を比較対象に選んだ理由は，ODP が Web ページを分類することを目的として作成されたディレクトリであり，ブックマーク管理サイトである del.icio.us と目的が似ているからである．本研究で抽出したタグの階層構造が ODP のディレクトリ構造に似ているならば，ナビゲーションとしての機能をもっていると考えられる．

3.5.2 評価手法

タグの階層構造 (以下タグ・ツリー) と ODP の階層構造 (以下 ODP ツリー) の間で，共通部分木構造の個数を数えることで類似度を評価する．共通部分木の数え上げには，3.2.1 節で紹介したラベル付き順序木間の畳み込みカーネルを利用した手法を用いた．ここで注意点として、タグ・ツリーや ODP ツリーはラベル付き順序木ではなく、ラベル付き順序なし木である．そのため，ここでは本来よりも厳しい条件で評価していることになる．

ラベル評価

式 (3.4) では木の頂点についているラベルを比較する関数 $L(v_1, v_2)$ が使われている．しかし，タグ・ツリーの各頂点に付けられているラベルと ODP ツリーに付けられているラベルでは，同じ意味を表していても文字列が異なる可能性が高い．そのため，単純にラベルの文字列を比較するだけでは不十分である．そこで，ラベルの評価は以下のように行う．

まず ODP のディレクトリ名は，例えば “News and Media”や “Data Formats”などのように複数の語からなるものが多い．そのためこれらを全て空白で区切り，単語のリストにする．次にこの単語のリストから “and”などのストップワードを取り除く．残った単語のリストをその ODP 頂点のラベルとして扱う．タグ頂点との比較では，ODP の単語リストのいずれかとタグ名が一致した場合に頂点同士が一致したものとみなす．なお，タグ名や ODP の単語リストは事前に全てステミング (stemming) 操作を行っておく．

また，シノニムに対応するために WordNet を利用する [32]．WordNet は英語の概念辞書で，英単語の品詞や単語の持つ意味概念が定義されている．また，単語間の関係を検索することもできる．タグ名と ODP の各単語を WordNet で検索し，シノニム関係と判定された場合には 2 つの頂点が一致したとする．

3.5.3 実験

タグ・ツリーと ODP ツリーの間の共通部分木の数を 3.2.1 節で述べたラベル付き順序木の畳み込みカーネルによって数える．タグ・ツリーには 3.4 節の実験で得た avg を用いる．

ODP ツリーとしては，ODP のトップページから四つの大きなディレクトリを選び，木構造を構築したものに対して実験を行う．今回の実験で使用した ODP ツリーを**表 3.3** に示す．タグ・ツリーと類似度の高そうなものとして “Computers”と “Science”という名前のディレクトリを選択した．これはタグ・ツリー内で

表 3.3: ODP ツリーの最大深さと要素数

Directory Name	depth	size
Computers	9	4144
Science	11	5428
Shopping	10	3269
Sports	10	7524

使われているタグにコンピュータや技術関連のものが多く見受けられたためである．逆に類似度が低そうなものとして “Shopping” や “Sports” といった ODP ツリーを選んだ．各ツリーの最大深さと要素数は表 3.3 に示す通りである．

類似度のスコアは式 (3.3) の木カーネルを用いて次の式で計算する．

$$\begin{aligned} Score(T_1, T_2) &= K(T_1, T_2) - \sum_{v_1 \in V(T_1)} K(T(v_1), T_2) \\ & where \quad T_1 \in tagtree, T_2 \in odptree \end{aligned} \tag{3.12}$$

ここで $T(v)$ は頂点 v だけからなる木構造を表す．木カーネル $K(T_1, T_2)$ では，頂点が一つしかない木でも部分木として扱うため，その一個の頂点のラベルさえ合えば共通部分木として数えられてしまう．今は木の構造の類似度を評価したいため，頂点が一つしかないような場合の値は減算する．

なお比較のため，タグ・ツリーに使われているタグの集合からランダムに木を構成して ODP ツリーとのスコアを計算する．ランダム・ツリーは要素数 2 から 86(avg の要素数の最大値) までの各要素数ごとにスコアを計算する．それぞれの要素数ごとに 5 つのランダム・ツリーを作成して評価を行った．

3.5.4 結果

まずタグ・ツリーと各 ODP ツリー間のスコアをまとめた結果を**図 3.12**，**図 3.13** に示す．図 3.12 は avg のタグ・ツリーのスコアだけをまとめた結果，図 3.13 はランダム・ツリーのスコアだけをまとめた結果である．

図より，全体としてタグ・ツリーの要素数が大きくなるとスコアもよくなる傾向があることがわかる．これはタグ・ツリーの要素数が増えるほど ODP ツリーと一致する頂点の数が増えるためであると考えられる．

また，タグ・ツリーとランダム・ツリーのどちらも “Computers”，“Science” とのスコアが高く，続いて “Shopping”，“Sports” の順にスコアが小さくなっていることがわかる．直感的に “Computers” や “Science” のほうがタグ・ツリーと類似した語彙が多かったため，これは予想されたとおりの結果である．この結果から，今回の評価手法が少なくともツリー間のラベルの類似度を測れることがわかる．

次に，ODP ツリーごとにタグ・ツリーとランダム・ツリーのスコアの比較を示した結果を**図 3.14**，**図 3.15**，**図 3.16**，**図 3.17** に示す．図 3.14∼ 図 3.17 を見る限り，タグ・ツリーとランダム・ツリーのスコアの間に有意な差は認められないことがわかる．これは，本研究の提案手法で得られた木と ODP のディレクトリ構造は似ていない，または本研究で行った木カーネルを利用した手法では評価できないという結果を示している．

3.5.5 考察

今回の評価では，提案手法で得られた木構造と ODP のディレクトリ構造は似ていない，または評価手法自体に評価する能力がないという結果になった．これらの可能性についてここでは考察を行う．

まずタグ・ツリーと ODP ツリーが似ていない可能性について検討してみる．これには，次の二つのことが考えられる．

- モデルが適切ではない

- タグ・ツリーと ODP ツリーの性質が元々異なる

まず我々の提案手法は以下のようなモデルに基づいている．

1. ユーザはオブジェクトが有している，あるいは表している概念をタグで表現して付加する
2. 様々な価値観を持つユーザが多くタグ付けすることで，様々な抽象レベルのタグがタグ付けされる
3. 広い意味のタグほど，多くのオブジェクトにタグ付けされやすい

これらの仮定が実際と大きく違う可能性がある．例えば上記の一番目の条件を考えると，実際にはオブジェクトの特性を表さないタグが付加されていることはめずらしくない．このようにモデルと現実との違いがタグ・ツリーに大きな影響を与えていると思われる．

次に，タグ・ツリーと ODP ツリーではその性質が元々異なるものである可能性が指摘できる．ODP ツリーは全体で一つの体系を成すように，あらかじめ考えられて作られている．対してタグ・ツリーはユーザ個人個人の持つ概念の階層構造を寄せ集めた形で作られていると言える．個人の持つ概念の階層構造は，体系的に作成された階層構造ほどには深くないことが予想される．そのため，体系的な階層構造との類似度が低い可能性がある．

以上を踏まえて上記の可能性の検証を行う必要がある．タグ・ツリーと ODP ツリーの性質の違いを検証することは恐らく難しいと考えられるため，モデルの正当性を検証を行うべきである．モデルの検証には，統計的な手法が利用できると考えられる．しかし，評価には最終的に人間の判断が介在してくることと予想され，定量的に評価を行うことは難しいと思われる．

次に評価手法について検討を行う．今回採用した評価手法では，ラベル付き順序木の共通部分木の数を数え上げることで類似度を測った．しかし実際にはタグ・ツリーや ODP ツリーは順序なし木であるので，本来より厳しい条件での評価であった．そのために，ランダム・ツリーとの差が出なかった可能性がある．今後はより緩い条件の木を比較できる手法を用いて評価する必要がある．なお，順序なし木の畳み込みカーネル関数の計算は NP 完全問題であることが証明されている [33]．

3.6 まとめ

本章では Folksonomy のタグの階層構造を構築する手法について述べた．また，得られた階層構造の評価として ODP のディレクトリ構造との類似度を計算することを試みた．

結果として，直感的には意味がありそうな階層構造を得ることができた．ただし，評価ではランダムに構築した階層構造との有意な差を得ることができなかった．

階層構造の評価に関しては，ツリーのラベルの類似度を評価できることが分かった．ただし，今回のようなタグ・ツリーと ODP ツリーの間の構造の類似度を評価することはできない可能性が示された．

また，3.5.5 節にて今回示された結果の考察を行った．今後はモデルの正当性の検証や，より緩い条件での評価手法などを行っていく必要がある．

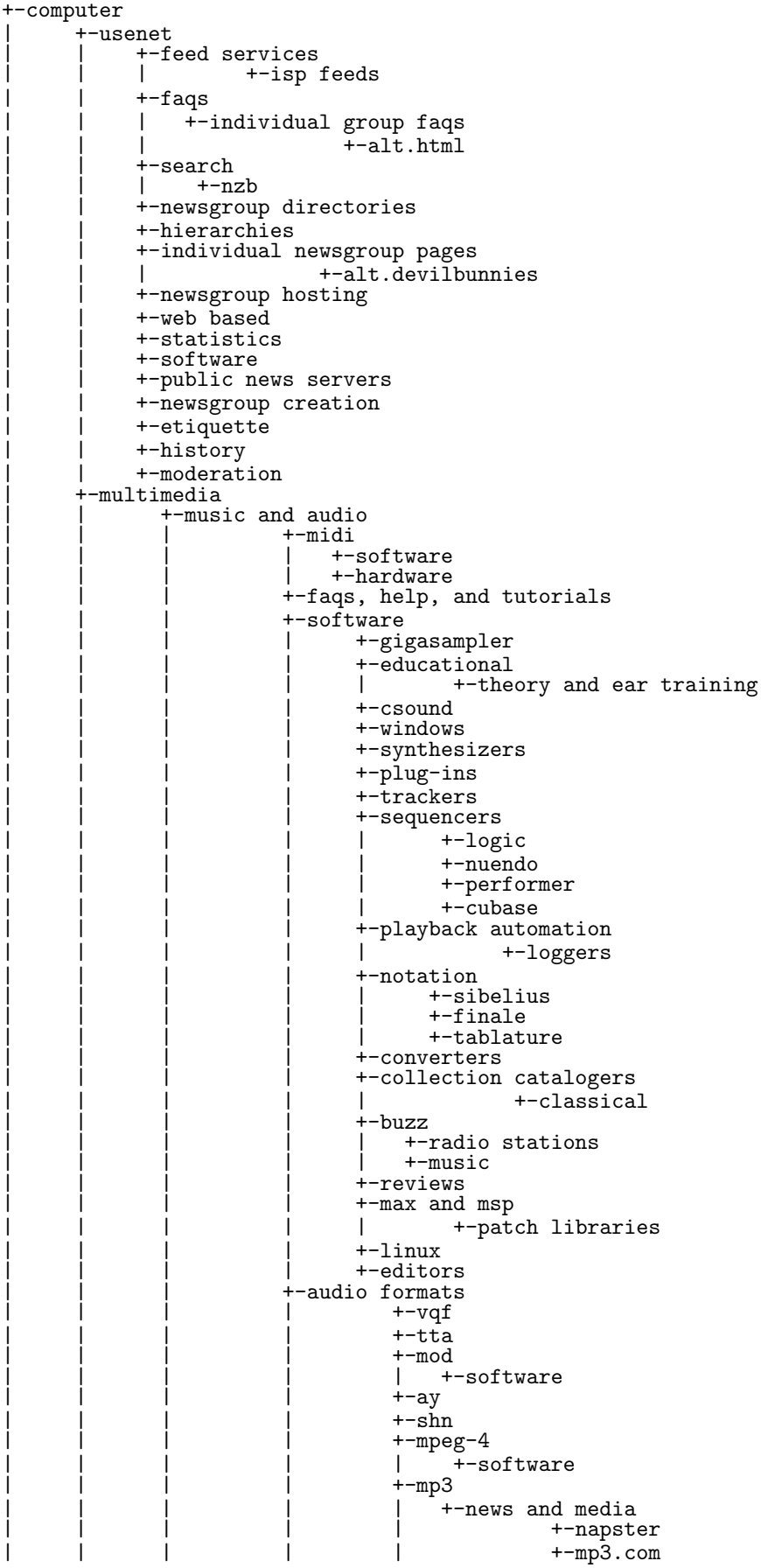


図 3.11: ODP の “Computers”ディレクトリの一部

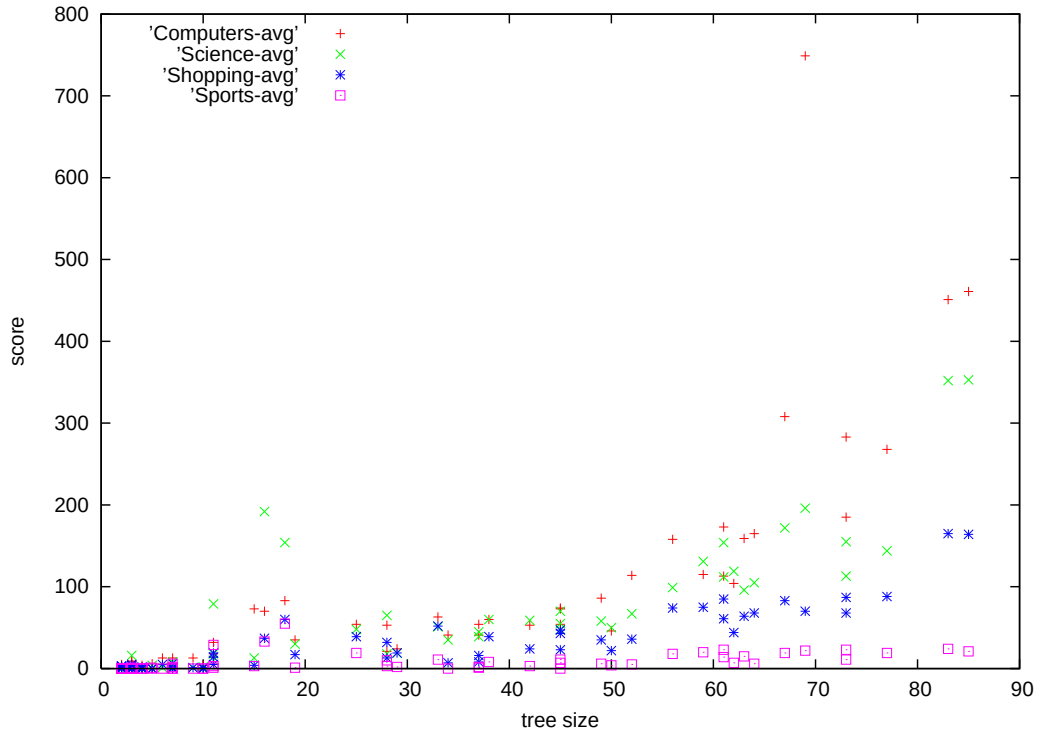


図 3.12: タグ・ツリーのスコア

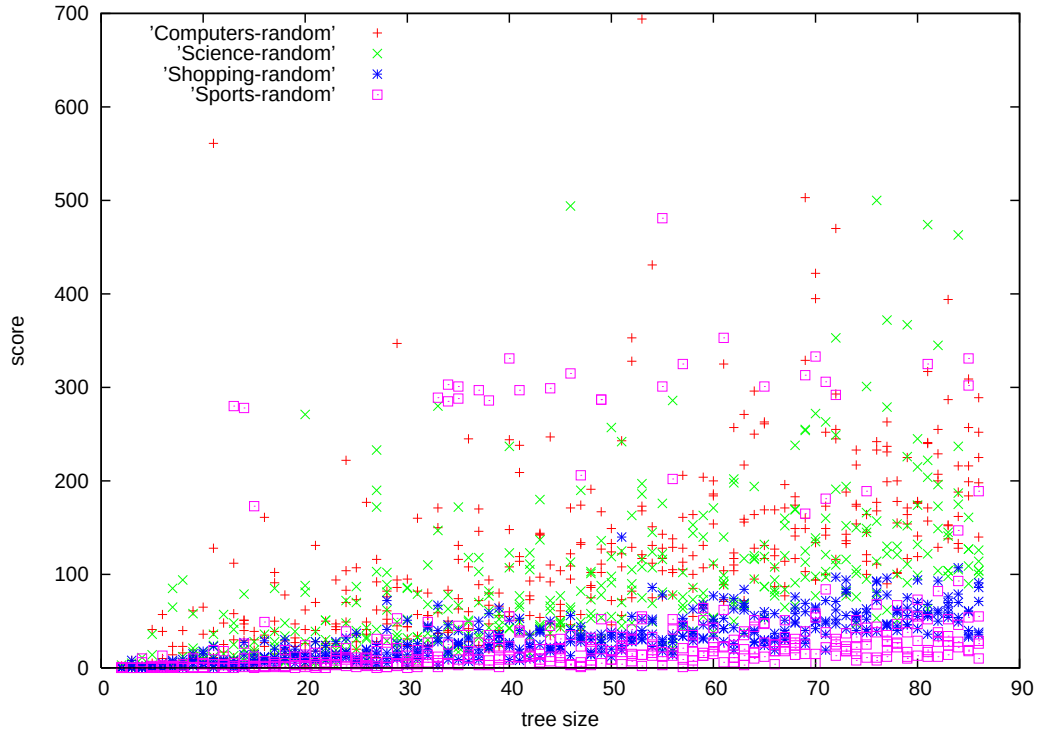


図 3.13: ランダム・ツリーのスコア

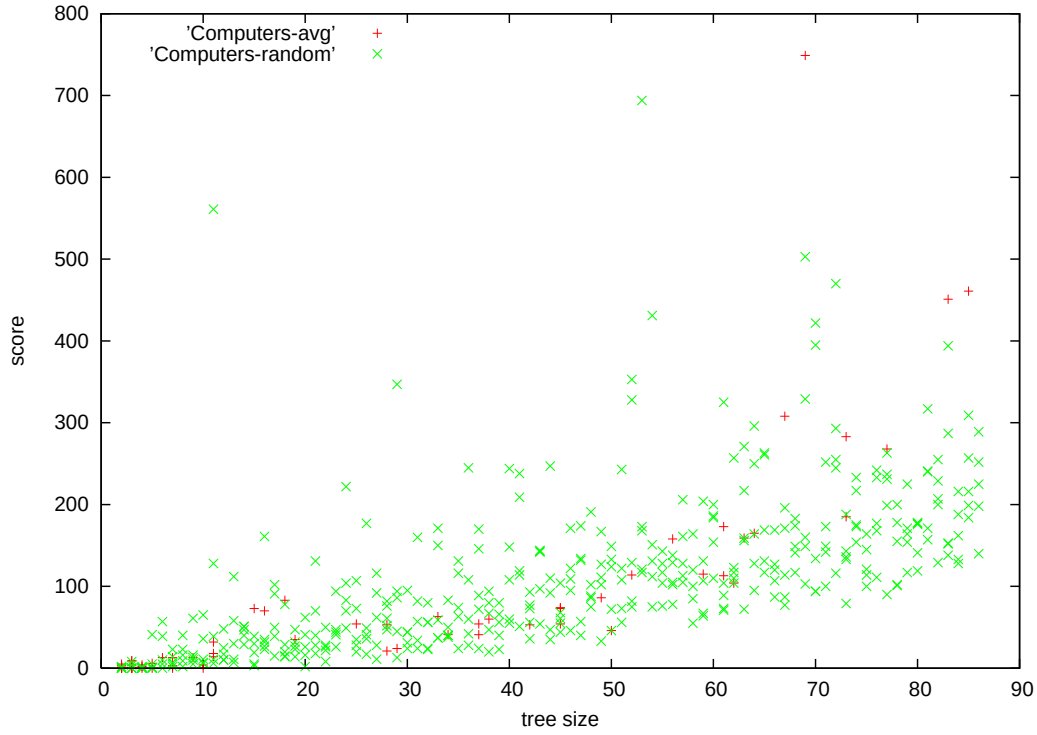


図 3.14: “Computers”との間のスコア

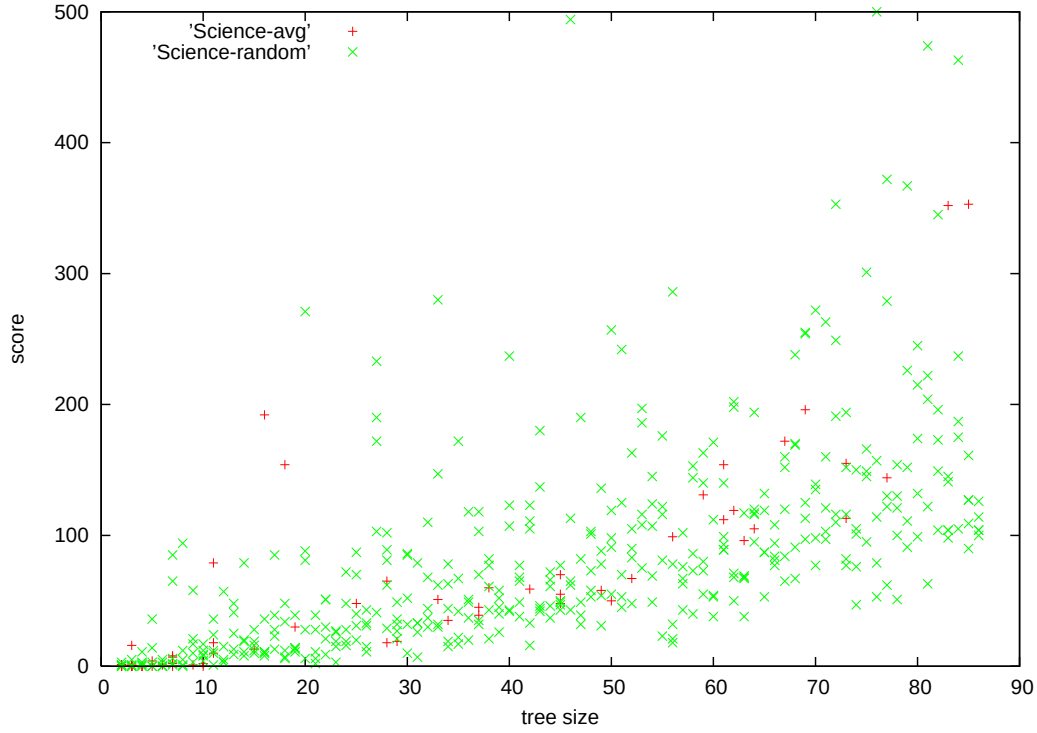


図 3.15: “Science”との間のスコア

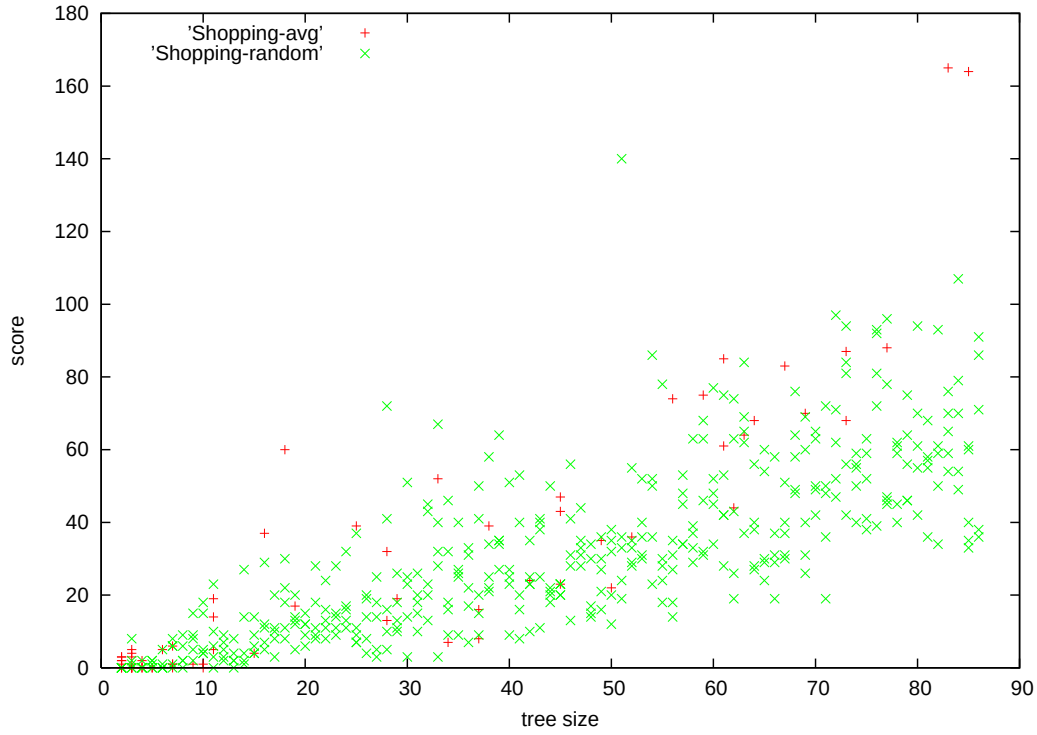


図 3.16: “Shopping”との間のスコア

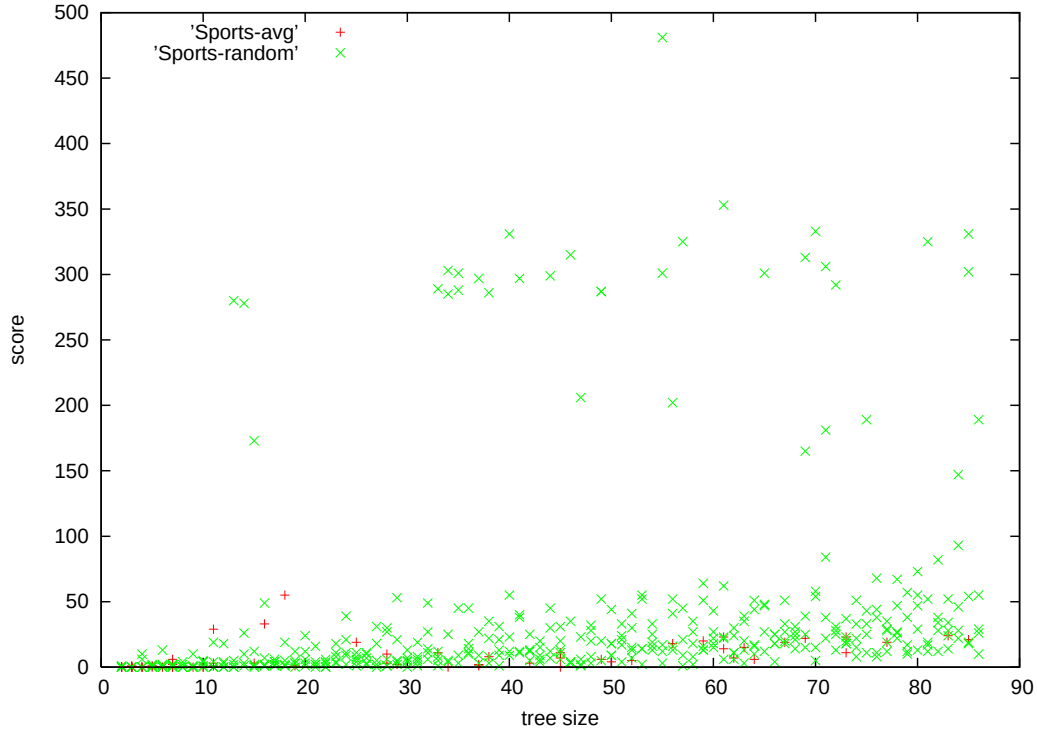


図 3.17: “Sports”との間のスコア

第4章　タグ付けの機械学習

4.1　背景

Folksonomy で大規模な数のオブジェクトを管理する場合の欠点の一つとして，オブジェクト全体の中でユーザがタグ付けした部分しか管理できない点が挙げられる．例えば Web ページの管理を行う del.icio.us の場合，Web 上にあるページを全て人手だけでタグ付けすることは難しい．大勢のユーザが集まってタグ付けするなどの対策は行っているが，マシンによるクローリングに比べると Folksonomy のシステムに蓄積される情報の収集量は小さいと言える．

そこで，Folksonomy でもマシンによるクローリングを行うことで情報の収集量を向上させることが考えられる．Folksonomy では情報の管理はタグで行うため，Web ページの内容から自動的に適切なタグをつける機能が必要となる．

本研究では機械学習の手法を使って Web ページから適切なタグを推測し，Folksonomy の情報収集量を向上させる手法を提案する．

4.2　関連研究

4.2.1　Support Vector Machine

Support Vector Machine(SVM) は Vaptik らによって提案された二値分類器の一つである [34]．SVM では超平面 (hyperplane) と特徴空間上の点との最短距離 (マージン) が最大化するように学習データを二分する．**図 4.1** に SVM の概要を示す．

学習

学習データの集合を

$$S = \{(\mathbf{x}_1, y_1), \ldots (\mathbf{x}_l, y_l)\}, \quad \mathbf{x}_i \in \mathbb{R}^N, y_i \in \{-1, 1\}$$

と表す．それぞれのデータは N 個の成分 (入力) とクラス分けのための指標 $\{-1, 1\}$(出力) からなっている．学習とは，これらのデータから識別関数 $f : \mathbb{R} \to \{-1, 1\}$ を選び，学習データにない未知のデータ $(\mathbf{x}, y)$ を正しく推定 $(f(\mathbf{x}) = y)$ することを目的とする．

線形 SVM

図のように学習データ集合 S が $\mathbb{R}^N$ の超平面によって $y_i = 1$ と $y_i = -1$ のクラスに分離できる場合は線形分離可能 (linearly separable) と言う．

候補となる超平面を

$$(\mathbf{w} \cdot \mathbf{x}) + b = 0 \tag{4.1}$$

と表す．ここで，$\mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R}$ で $\mathbf{w}$ は超平面の法線ベクトルとする．そして，識別関数を

$$f(x) = \mathrm{sgn}((\mathbf{w} \cdot \mathbf{x}) + b) \tag{4.2}$$

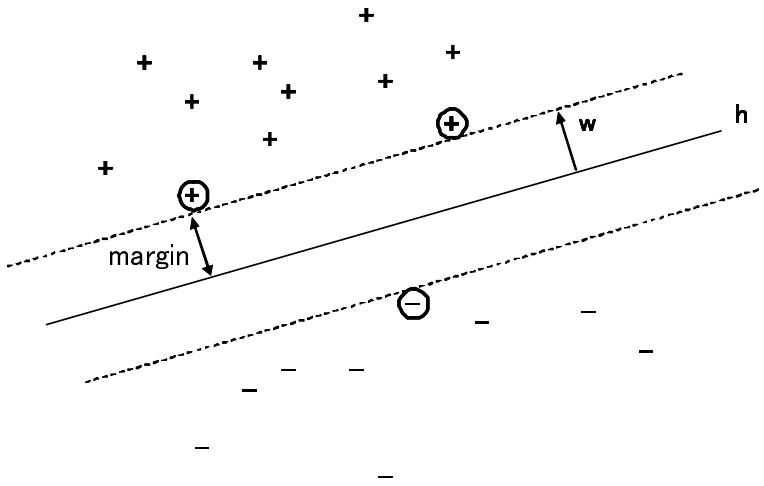


図 4.1: Support Vector Machine の概要

とする．ただし，ここでは $\mathrm{sgn}(0) = 1$ とする．

超平面 $(\mathbf{w}, b)$ に対するサンプル $(\mathbf{x}_i, y_i)$ のマージンは

$$\gamma_i = \frac{y_i((\mathbf{w} \cdot \mathbf{x}_i) + b)}{||\mathbf{w}||} \tag{4.3}$$

と定義される．$\gamma_i > 0$ ならば $(\mathbf{x}_i, y_i)$ が正しく識別されていることになる．

学習データ S が線形分離可能であった場合，

$$y_i((\mathbf{w} \cdot \mathbf{x}_i) + b) > 0, \quad i = 1 \ldots l$$

となるので，超平面を正規化して

$$\min_i \{y_i((\mathbf{w} \cdot \mathbf{x}_i) + b)\} = 1 \tag{4.4}$$

と表すことができる．この形式を正準形 (canonical form) と言う．このときマージンの長さは $1/||\mathbf{w}||$ になるので，マージンを最大化するには $||\mathbf{w}||$ を最小化すればよい．そのため，マージンが最大となるような超平面は，

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}||\mathbf{w}||^2 \\ &\text{subject to} \quad y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, \quad i = 1, \ldots l \end{aligned} \tag{4.5}$$

を解くことによって得られる．ここでラグランジュの未定乗数法を用いることにより，以下の双対問題へと変換できる．

$$\begin{aligned} &\text{maximize} \quad L(\mathbf{w}, b, \boldsymbol{\alpha}), \\ &\text{subject to} \quad \nabla_{\mathbf{w}} L(\mathbf{w}, b, \boldsymbol{\alpha}) = 0, \nabla_b L(\mathbf{w}, b, \boldsymbol{\alpha}) = 0, \boldsymbol{\alpha} \geq \mathbf{0} \\ &\text{where} \quad L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}||\mathbf{w}||^2 - \sum_{i=1}^{l} \alpha_i (y_i((\mathbf{w} \cdot \mathbf{x}_i) + b) - 1), \quad \mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R}, \boldsymbol{\alpha} \in \mathbb{R}^l \end{aligned}$$

このとき最適性条件である Karush-Kuhn-Tucker 条件は

$$\nabla_{\mathbf{w}} L(\mathbf{w}, b, \boldsymbol{\alpha}) = \mathbf{w} - \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i = 0 \tag{4.6}$$

$$\nabla_b L(\mathbf{w}, b, \boldsymbol{\alpha}) = \sum_{i=1}^{l} \alpha_i y_i = 0 \tag{4.7}$$

$$\alpha_i (y_i((\mathbf{w} \cdot \mathbf{x}_i + b) - 1)) = 0, \quad \alpha_i \geq 0, \quad y_i((\mathbf{w} \cdot \mathbf{x}_i) + b) - 1 \geq 0, \quad i = 1, \ldots l \tag{4.8}$$

これを解いていくと双対問題は

$$\begin{aligned} \text{maximize} \quad & -\frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) + \sum_{i=1}^{l} \alpha_i, \quad \boldsymbol{\alpha} \in \mathbb{R}^l \\ \text{subject to} \quad & \sum_{i=1}^{l} \alpha_i y_i = 0, \boldsymbol{\alpha} \geq \mathbf{0} \end{aligned} \tag{4.9}$$

となる．

さて式 (4.8) の条件より，非零の α_i に対応する学習データ $(\mathbf{x}_i, y_i)$ は以下を満たす．

$$y_i((\mathbf{w} \cdot \mathbf{x}_i) + b) = 1$$

式 (4.4) よりこれらの点は超平面に最も近い点の集合である．式 (4.6) より，式 (4.9) の解を $\boldsymbol{\alpha}^*$ とすると最適な超平面 $(\mathbf{w}^*, b^*)$ は非零の α_i^* に対応する学習データから以下のように表される．

$$\mathbf{w}^* = \sum_{\mathbf{x}_i \in SV} \alpha_i^* y_i \mathbf{x}_i \tag{4.10}$$

$$b^* = y_i - \mathbf{w}^* \cdot \mathbf{x}_i, \quad \mathbf{x}_i \in SV \tag{4.11}$$

ここで，超平面を構成するデータをサポートベクタ (Support Vector) と呼び，SV と表す．

以上により分類器の識別関数は

$$\begin{aligned} f(x) &= \text{sgn}(\mathbf{w}^* \cdot \mathbf{x} + b^*) \\ &= \text{sgn}\left(\sum_{\mathbf{x}_i \in SV} \alpha_i^* y_i (\mathbf{x}_i \cdot \mathbf{x}) + b^* \right) \end{aligned} \tag{4.12}$$

と求めることができる．式 (4.12) に見るように，識別関数の計算にはサポートベクタのみを用いて計算すればよい．一般にサポートベクタは学習データに対して非常に小さくなることが多いため計算量を削減することができる．

ソフトマージン最適化

前節では線形分離可能な場合を説明したが，一般に多くの問題は線形分離不可能であることが多い．そのような場合には，元の式 (4.5) の問題にスラック変数を導入し以下のように変換する．

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}||\mathbf{w}||^2 + C \sum_{i=1}^{l} \xi_i \\ \text{subject to} \quad & y_i((\mathbf{w} \cdot \mathbf{x}_i) + b) \geq 1 - \xi_i, \xi_i \geq 0 \quad i = 1, \ldots l \end{aligned} \tag{4.13}$$

ここで，$C > 0$ は識別誤りに対するペナルティパラメータである．このような形式の制約条件を扱うものをソフトマージンによる最適化という．**図 4.2** にその概要を示す．式 (4.13) は式 (4.5) と同様に双対問題から

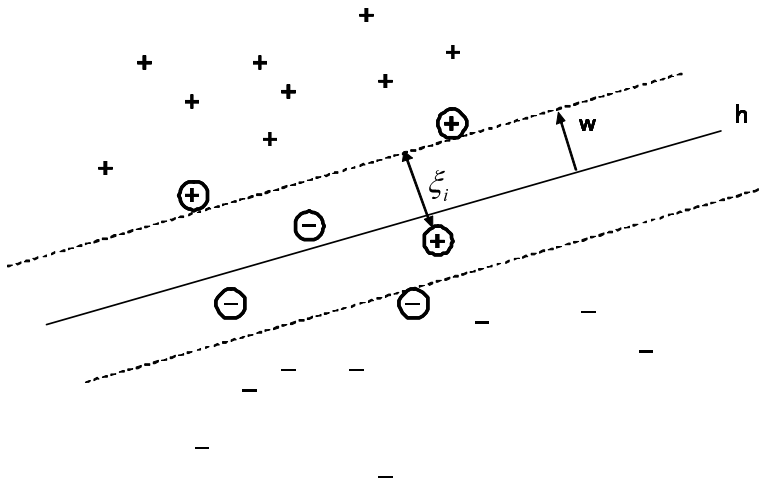


図 4.2: ソフトマージン最適化の概要

解くことができる．解は以下で表される．

$$\begin{aligned} \mathbf{w}^* &= \sum_{\mathbf{x}_i \in SV'} \alpha_i^* y_i \mathbf{x}_i \\ b^* &= y_i - (\mathbf{w}^* \cdot \mathbf{x}_i), \quad \mathbf{x}_i \in SV \\ f(x) &= \mathrm{sgn}\left(\sum_{x_i \in SV'} \alpha_i^* y_i (\mathbf{x}_i \cdot \mathbf{x}) + b^* \right) \\ & \quad \text{subject to} \quad 0 \leq \alpha_i^* \leq C \end{aligned} \tag{4.14}$$

ここで，SV は識別誤りではない通常のサポートベクターの集合，SV' は識別誤りのものも含んだサポートベクターの集合を表す．

ソフトマージンのアルゴリズムを使うことで多少の識別誤りを吸収しつつ学習を行うことができる．

4.2.2 自動的なタグ付けの研究

Web ページに自動的にメタデータを付加する研究は，セマンティック Web の分野で広く行われている．2.4.2 節で述べたようにセマンティック Web ではオントロジーで定式化された概念を表すメタデータをあらかじめ Web ページに付加 (アノテーション) することが必要となる．アノテーションを人手でやるのは負担が大きいため，自動的にアノテーションを行う手法が提案されている [35, 36, 37]．特に Heßらは機械学習を利用して自動的にアノテーションを行う手法を提案している [38, 39]．しかし精度は低く，あくまで人間がアノテーションを行う際の手助けを前提としている．

オントロジーではなくタグを自動的に付加する研究も近年行われている．Chirita らは，各ユーザが自身のパソコンに保存しているファイルと，Web 上の文書の間の類似度を計算し，似ているファイルのキーワードをタグとして付加する手法を提案している [40]．この手法は各個人のファイル内のキーワードからタグを生成するため，個人化 (personalized) されたタグを付加することができる．

4.3 提案手法

4.3.1 概要

図 4.3 に提案手法の概要を示す．本手法ではまず Folksonomy 上ですでにタグが付加されている Web 文書

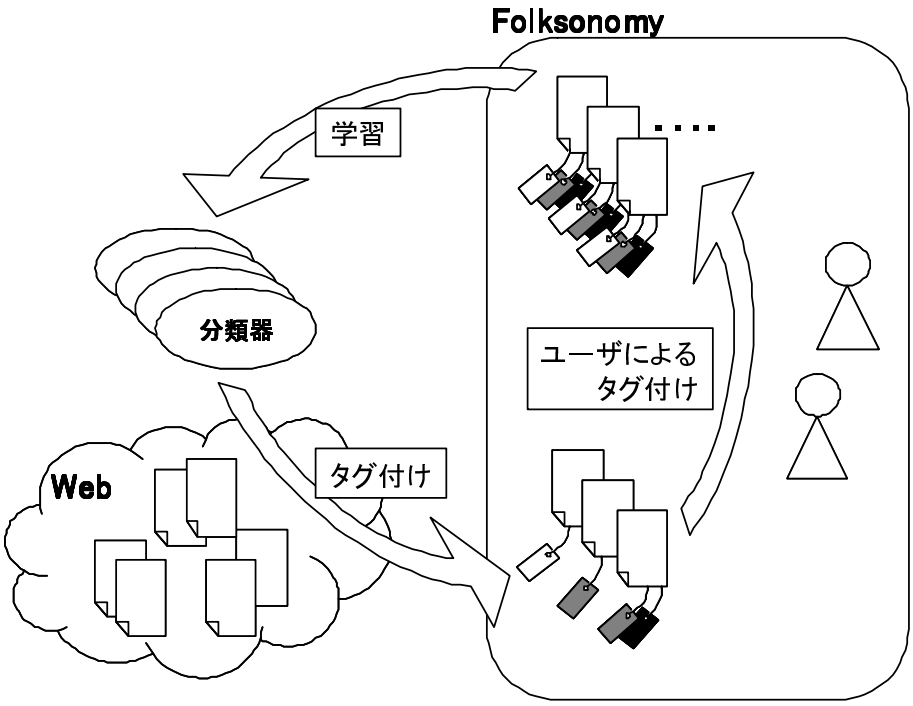


図 4.3: タグ付けの分類器を用いた Web クローリングの概要

を学習データとしてタグ付けの学習を行い，タグを付加すべきかどうかを判定する分類器を作成する．分類器は各タグにつき一つ作成する．次に，作成した分類器を用いて未知の Web 文書にタグを付加する．タグ付けすることによって Folksonomy 上でこの文書を管理することができ，ユーザがその文書を検索して探すこともできるようになる．ユーザの目に留まった文書はユーザによってタグ付けされていくことで新たな学習データとなる．

本研究ではタグ付けの学習に線形の SVM を用いる．これは以下の理由による．

- 高速に分類を行うことが出来る
- テキスト分類の分野で良い性能を示している

SVM はあらかじめ学習を済ませておくことで，式 (4.12) に見るようにベクトルの内積を計算すれば分類することができる．クローリングを行うにはできるだけ処理は高速に行ったほうが都合が良い．

また，SVM はテキスト分類の分野で良い性能を示すことが知られている [41]．本研究で行うタグ付けの学習はテキスト分類と同様に考えることができると思われる．そのため，テキスト分類で良い性能を示している SVM は効果を期待できる．また，テキスト分類は多くの場合線形分離可能であることが経験的に分かっている [42]．そのため本研究では線形の SVM を用いて実験を行った．また，線形の SVM を用いることで一つ目の理由である高速化の効果もある．

4.3.2 タグの選択

本手法では各タグに対し，そのタグが付加されるかどうかを判定する分類器を作成する．しかし，全てのタグに関して分類器を作るのは以下の点で不都合である．まずタグはそれぞれの利用頻度に大きな差があるため，利用頻度の低いタグの分類器までも作成するのは非効率と言える．また全てのタグの分類器を作成することは計算にかかる処理も大きくなってしまう．

さらに，より重要な問題として，タグと文書の内容が合っていないなどといった「雑音」が学習精度の低い分類器によって多く生産されてしまう．

以上の理由から，分類器を作成するタグを選択する必要がある．以下では分類器を作成するタグをどのようにして選ぶかについて説明する．

本研究では以下の指標を用いて分類器を作成するタグを選択する．

$$r_t = \frac{n_t}{s_t} \tag{4.15}$$

ここで n_t はタグ t の総タグ付け数，s_t は t のサイズを表す．タグのサイズとはすなわちタグ付けされているオブジェクトの数なので，式 (4.15) は一つのオブジェクトに付加されているタグ t の平均数と見ることができる．

式 (4.15) を指標に用いたのは以下の仮定による．すなわち一つのオブジェクトにタグ付けされているタグの数を，ユーザからの投票と見なす．すると，多くの投票を得たタグというのは，多くのユーザがオブジェクトを見て，そのタグを付加することに同意したものであるということができる．すなわち文書中にそのタグに関する特長が強く出ていると考えられる．

本研究では式 (4.15) の指標に応じてタグの分類器を作成し，その精度の評価を行った．

4.4 実験方法

本研究では，各タグごとに対応する分類器を一つ作成する．あるタグ t に対応する分類器は，文書の特徴を入力としてタグ t がつくかどうかを判定する．本節では本研究で行った SVM の学習の基本設定について述べる．

4.4.1 特徴ベクトルの抽出

本研究では文書の特徴は，その文書に出現する単語によって表す．各文書から以下の手順に従って特徴ベクトルを抽出する．

事前処理

各文書から単語を抽出する．この時 HTML タグや javascript のプログラムなど，文書の内容と直接関係のないものは排除する．またストップワードの除去も行う．抽出した各単語は Porter Stemming [43] のアルゴリズムを使って語幹抽出を行っておく．

次元削減

抽出した単語全体の中から，実際に特徴として用いる単語を選択する．本研究では以下の情報利得を基に単語の選択を行った．

$$\begin{aligned} G(t) &= -\sum_{i=1}^{n} P(c_i)\log P(c_i) \\ &\quad +P(t)\sum_{i=1}^{n} P(c_i|t)\log P(c_i|t) \\ &\quad +P(\bar{t})\sum_{i=1}^{n} P(c_i|\bar{t})\log P(c_i|\bar{t}) \end{aligned} \tag{4.16}$$

ここで n はカテゴリ数であり，$P(c_i)$ はある文書がカテゴリ c_i に属する確率を表す．また t は単語であり，$P(t)$ は文書中に t が出現する確率である．式 (4.16) はカテゴリ分類における単語 t の持つ情報量を表現する．本研究では情報利得の値が高い単語から順に特徴として選択した．

単語の重み付け

文書の特徴をより表していると思われる単語に対して重みを与える．ここではテキスト分類で一般に用いられている tf·idf [44] を用いて単語の重み付けを行う．tf·idf は以下の式で定義される．

$$\mathrm{tf} \cdot \mathrm{idf}(t_i, d_j) = tf_{t_i,d_j} \cdot \log \frac{N}{df_{t_i}} \tag{4.17}$$

ここで，tf_{t_i,d_j} は文書 d_j に単語 t_i が出現する回数，df_{t_i} は t_i が出現する文書の数，N が文書の総数を表す．これにより文書中に多く出現し，かつ他の文書にはあまり出現しないような単語に大きな重みを与えることができる．実際には以下に示すように正規化した値を用いて重み付けを行った．

$$\frac{\mathrm{tf} \cdot \mathrm{idf}(t_i, d_j)}{\sqrt{\sum_k (\mathrm{tf} \cdot \mathrm{idf}(t_k, d_j))^2}} \tag{4.18}$$

4.4.2 学習データ

学習データとして，タグ付きの Web 文書 30,000 を次のようにして用意した．2007 年 2 月から 2007 年 12 月の期間に del.icio.us からクローリングして得た全 433,540 のブックマークされた文書の中からランダムに 30,000 を選択したものを用いる．ただし，今回の実験ではテキストデータのみを対象とし，動画や画像などに直接ブックマークされたものは選択肢から除外した．なお，文書に使われている言語は英語が圧倒的に多かったため，今回の実験では言語によるフィルタリングは行っていない．

学習データのクラス分け

学習データの文書について，あるタグ t がタグ付けされているものを正例，タグ付けされていないものを負例として二つのカテゴリに分類する．本研究ではタグ t が閾値 5 以上タグ付けされていた場合を正例とした．これは誤ってつけられたタグなどの雑音の影響を軽減するためである．

実験に使用するタグ

本実験で学習を行うタグを次のようにして選んだ．まずタグの全集合からサイズの値が大きい順に 100 のタグを選んだ．次に，選んだ 100 のタグに対して式 (4.15) の指標を計算し，その値の上位 20 のタグ (1 位 ~20 位) と下位 20 のタグ (81 位 ~100 位) を実験に使用した．

サイズの値が大きいタグを最初に選んだ理由は，正例となる学習データの割合を増やすことで，正例が足りずに学習が十分にできないという状況を回避するためである．

4.4.3 分類器の評価方法

作成した分類器の評価方法としては，5-fold cross validation を用いる．N-fold cross validation は，まず学習データを N 個の組に等分割してそのうちの N-1 組のデータを用いて分類器の学習を行い，その学習で得た分類器で残り 1 組の分類を行う．これを全 N 通りの組み合わせ全てに関して行う．

4.4.4 実験環境

SVM にライブラリとして提供されている LIBLINEAR [45] を使用した．これは，同様に公開されている LIBSVM [46] を線形専用に改良を施したものである．

表 4.1: 上位 20 タグの分類器の性能

n	tag name	precision	recall	f-score
1	css	0.843	0.634	0.724
2	javascript	0.848	0.609	0.709
3	ajax	0.797	0.494	0.610
4	google	0.860	0.528	0.654
5	design	0.786	0.637	0.703
6	photography	0.817	0.607	0.697
7	webdesign	0.793	0.653	0.716
8	mac	0.870	0.598	0.709
9	flash	0.800	0.415	0.547
10	music	0.834	0.562	0.672
11	linux	0.832	0.621	0.711
12	web	0.711	0.502	0.589
13	productivity	0.739	0.420	0.536
14	programming	0.802	0.703	0.749
15	search	0.694	0.358	0.472
16	security	0.794	0.536	0.640
17	tools	0.686	0.601	0.641
18	java	0.831	0.486	0.613
19	games	0.857	0.524	0.650
20	software	0.748	0.662	0.702
microavg.		0.772	0.594	0.671

表 4.2: 下位 20 タグの分類器の性能

n	tag name	precision	recall	f-score
81	culture	0.570	0.347	0.431
82	utilities	0.621	0.285	0.391
83	hacks	0.626	0.267	0.374
84	internet	0.598	0.324	0.420
85	download	0.624	0.315	0.419
86	media	0.599	0.324	0.421
87	technology	0.585	0.301	0.398
88	code	0.579	0.367	0.449
89	online	0.602	0.258	0.361
90	cool	0.558	0.306	0.395
91	resources	0.570	0.231	0.328
92	research	0.595	0.307	0.405
93	computer	0.599	0.274	0.376
94	tool	0.549	0.275	0.367
95	tech	0.570	0.226	0.323
96	interesting	0.495	0.238	0.321
97	article	0.591	0.396	0.474
98	system	0.584	0.467	0.519
99	toread	0.607	0.392	0.476
100	imported	0.483	0.144	0.222
microavg.		0.581	0.320	0.412

4.5 実験結果

4.5.1 結果

表 4.1 と**表 4.2** に各タグごとの分類器の性能の結果を示す．表の結果は特徴ベクトルの次元数を 10000 までに削減して計算したものである．表 4.1 は上位 20 のタグ (以下 top20)，表 4.2 は下位 20 のタグ (以下 bottom20) の結果を表している．

表 4.1 と表 4.2 の結果を比べると，明らかに top20 の分類器の方が bottom20 の分類器よりも性能が良いことが分かる．それぞれの評価値のマイクロ平均をとると，top20 の各評価値は bottom20 のタグの評価値に比べて precision で約 0.2，recall で約 0.27，f-score で約 0.26 の増加となっている．

また，**図 4.4・図 4.5・図 4.6** に特徴ベクトルの次元数を変化させた場合の top20，bottom20 それぞれの性能を示す．各図の値は top20，bottom20 それぞれでマイクロ平均をとった結果である．図 4.4〜 図 4.6 より，各評価値の全ての次元において top20 が bottom20 よりも良い値を示していることが分かる．

これらの結果から，式 (4.15) の指標がタグの選択に有効であることがわかる．

4.5.2 考察

ここでは 4.5.1 節で得られた結果から，タグの分類器を用いて Folksonomy におけるクローリングを行うことについての考察を行う．

まず特徴ベクトルの適切な次元数について考える．図 4.4〜 図 4.6 より，各評価値ごとに次元数に対する振る舞いが異なることが分かる．そのため，何を重要視するかで適切な次元数が異なってくる．図の結果より，

表 4.3: 次元数 1000 と 10000 における top20 の性能比較

次元数	precision	recall	f-score
1000	0.782	0.559	0.652
10000	0.772	0.594	0.671

次の二つの選択肢が考えられる．

- precision を重視
  図 4.4 より，precision は次元数が小さい方が値が高いことが分かる．top20 の precision のピークは次元数が 1000 の時である．
- 総合的な性能重視
  f-measure は分類器の総合的な性能を示すと考えられる．図 4.6 によると，top20 では次元数が 10000 の時に値が横ばいとなる．

なお，recall の値がそこまで大きくないこと，f-measure の場合と同じく次元数が 10000 の辺りで recall の値が横ばいとなることから，recall を重視するという選択肢は除いた．

表 4.3 に次元数 1000 と 10000 のときの分類器の性能を比較したものを示す．どちらも top20 の各指標のマイクロ平均を取ったものである．次元数が 1000 の時の方が，10000 の時に比べて precision が 1%大きい．しかし，recall は 10000 の方が 3.5%大きく，f-measure も 1.9%大きい．一方，1000 の方は次元数が小さいので分類を 10000 よりも速く処理することが可能である．

このように，どちらにもメリットデメリットがあるため，どちらが良いかということははっきりと言えない．実際の要求に合わせて選択するのが良いと思われる．

4.6 まとめ

本章では Folksonomy におけるタグ付けを SVM を用いて学習することで，自動的に Web 文書にタグを付加する分類器を作成した．また，オブジェクト当たりのタグ付け数を指標として用いて性能の良い分類器を選ぶ方法を提案した．

結果として，作成した分類器は特徴ベクトルの次元数を 10000 とした時に precision が 0.772，recall が 0.594，f-measure が 0.652 の性能を得た．また，precision を重視する場合には特徴ベクトルの次元数は 1000 程度と小さく抑えられることがわかった．

さらに，提案した指標に関しては，性能の良い分類器を選択するために有効であることが示せた．

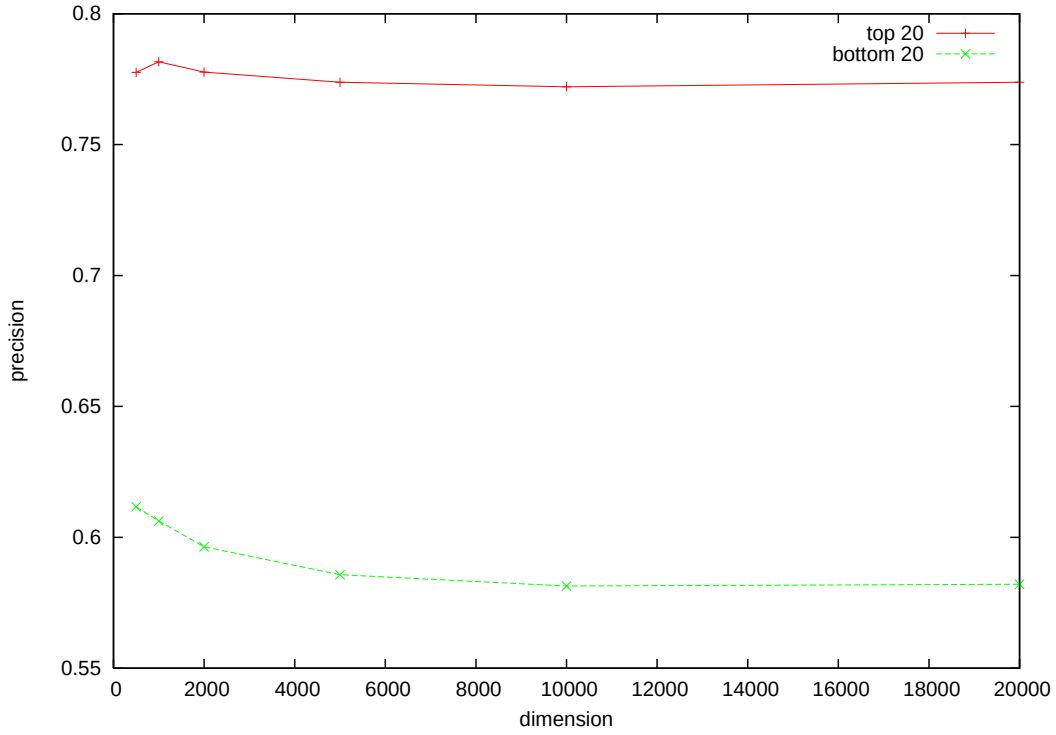


図 4.4: 次元数ごとの分類器の precision

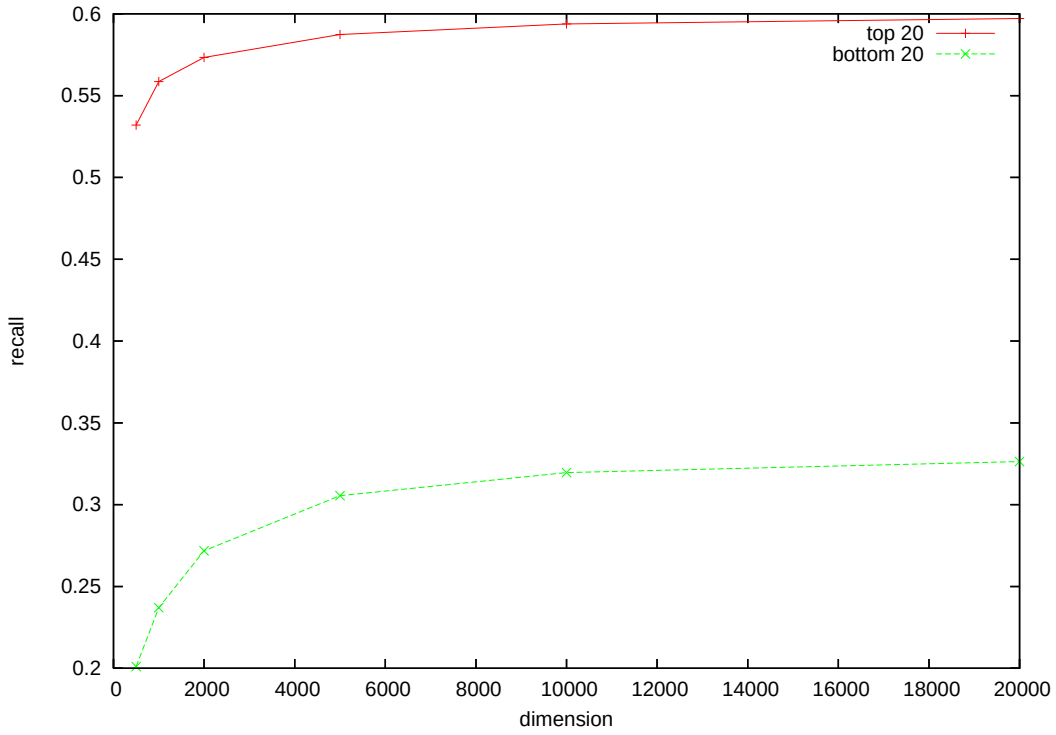


図 4.5: 次元数ごとの分類器の recall

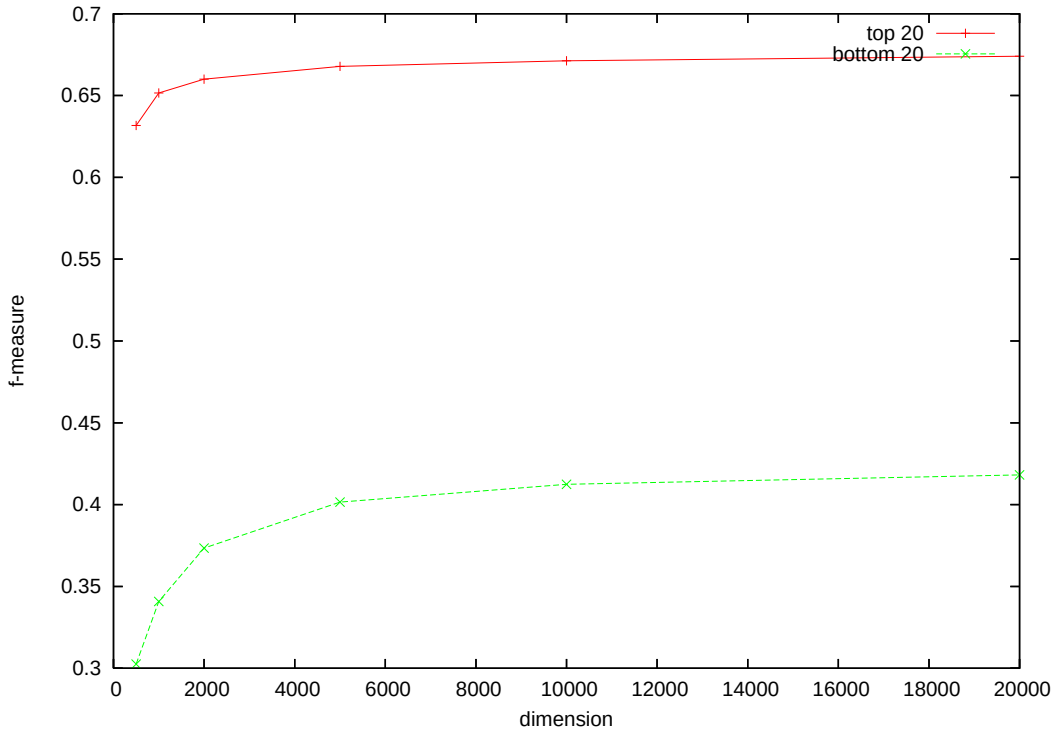


図 4.6: 次元数ごとの分類器の f-measure

第5章　結論

5.1　まとめ

本研究では Folksonomy を対象とした新たな Web マイニング手法を二つ提案した．一つ目としてタグの階層構造抽出を，二つ目として SVM を利用したタグ付けの機械学習を行った．また，階層構造を評価する手法として ODP のディレクトリ構造と共通部分木の数をかぞえて類似度を測る方法を試みた．

階層構造の抽出の結果としては，直感的には有用な階層構造を得ることができたと思われる．しかし，評価の結果としてはランダムに構築した階層構造との有意な差を得ることはできなかった．

階層構造の評価に用いた手法は，3.5.4 節で示した通り，タグ・ツリーと ODP ツリー間のラベルの類似度を評価する能力は確かめられた．しかし，構造の類似度を評価する能力はない可能性が示された．

SVM を利用したタグ付けの機械学習においては，4.5.1 節で示した通り，特徴ベクトルの次元数を 10000 とした時に，マイクロ平均で precision が 0.772，recall が 0.594，f-measure が 0.652 という値を得た．recall はそれほど高くないが，自動的なタグ付けで重要となる precision の値がそれなりに高いため，有効な分類器をつくることができたといえる．また，precision を重視する場合にはさらに次元数を削減できることが分かった．

なお，性能の良いタグの分類器の選択には，式 (4.15) で示した通り，オブジェクト当たりのタグ付け数を指標として用いるとよいことがわかった．

5.2　今後の課題

今後の課題としては以下の点が挙げられる．

- 階層構造抽出に関して
  - 階層構造構築の提案モデルの検証
    提案しているモデルが，どれくらい現実に沿っているかの検証を今回行うことができなかった．タグ付けの統計的な処理からある程度の検証は行えると期待できる．
  - 階層構造の新たな評価手法の検討
    今回評価に用いた手法は，ODP ディレクトリとの類似度を測るものであった．しかし，条件が厳しすぎて適切な評価が行われなかった可能性がある．そのため，より緩い条件で評価できる手法を検討すべきである．また，ODP ディレクトリと比較する以外に階層構造を評価する手段を検討することも必要である．
- タグ付けの機械学習に関して
  - SVM 以外の学習器の利用
    今回はテキスト分類で良い成績を収めている等の理由で SVM を利用したが，タグ付けの学習とテキスト分類では性質が異なる可能性がある．そのため，他の学習器を用いて分類器の性能を比較することが望ましい．
  - Folksonomy の情報をもっと活用する
    今回の手法では一つ一つのタグを独立に扱った．しかし，実際には一つのオブジェクトには複数

のタグが分布している．そのようなより豊富な情報を利用することで学習の精度を向上させることが可能ではないかと考えられる．

Appendix

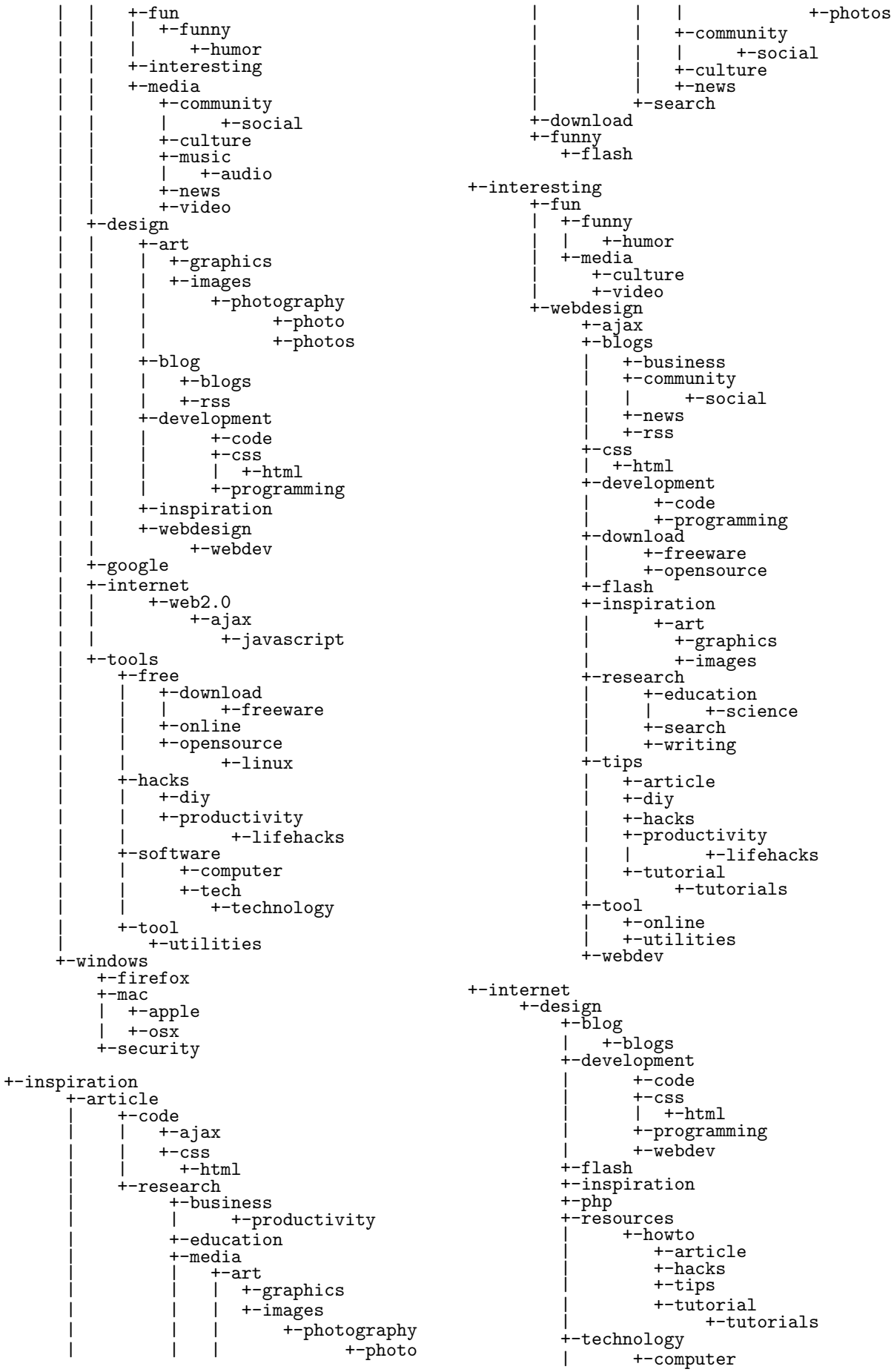
タグの階層構造 (avg)

```
+-ajax
   +-css
   |  +-html
   |  +-javascript
   |  +-library
   |  +-php
   +-social

+-art
  +-community
  |     +-culture
  +-funny
  +-graphics
       +-flash
       +-images
       |    +-photography
       |          +-photo
       |          +-photos
       +-photoshop

+-article
    +-code
    |   +-css
    |     +-html
    +-hacks
    +-research
         +-business
         |     +-productivity
         |           +-lifehacks
         +-education
         +-news

+-blog
   +-technology
         +-blogs
         |   +-rss
         +-computer
         +-firefox
         +-flash
         +-google
         +-interesting
         |      +-fun
         |      |  +-funny
         |      |     +-humor
         |      +-inspiration
         |      |      +-art
         |      |        +-graphics
         |      |        +-images
         |      |            +-photography
         |      |                  +-photo
         |      |                  +-photos
         |      +-writing
         +-mac
         +-media
         |   +-community
         |   |     +-social
         |   +-culture
         |   +-news
         |   +-video
         +-research
         |     +-business
         |     +-education
         |     +-search
         +-resources
         |     +-free
         |     |   +-download
         |     |   +-freeware
         |     |   +-online
         |     |   +-opensource
         |     |   +-tool
         |     |      +-utilities
         |     +-howto
         |     |   +-article
         |     |   +-diy
         |     |   +-hacks
         |     |   +-productivity
         |     |   |       +-lifehacks
         |     |   +-tips
         |     |   +-tutorial
         |     |       +-tutorials
         |     +-library
         +-web2.0
             +-ajax
             |   +-javascript
             +-webdesign
                  +-css
                  |  +-html
                  +-development
                  |      +-code
                  |      +-programming
                  +-webdev

+-blogs
   +-inspiration
   |      +-art
   |      |  +-graphics
   |      +-research
   |           +-business
   |           +-education
   |           +-media
   |           |   +-community
   |           |   |     +-social
   |           |   +-culture
   |           |   +-news
   |           |   |   +-rss
   |           |   +-video
   |           +-search
   |           +-writing
   +-tutorials
        +-article
        +-code
        |   +-ajax
        |   +-css
        |     +-html
        +-hacks
        +-utilities
             +-download
             |     +-opensource
             +-productivity
                   +-lifehacks

+-business
```

```
     +-education
     +-utilities
          +-ajax
          +-productivity
          |       +-lifehacks
          +-search
              +-community
              |     +-social
              +-news
                 +-culture

+-code
   +-opensource
        +-ajax
        |   +-css
        |   |  +-html
        |   +-java
        |   +-javascript
        |   +-php
        +-freeware
        +-library

+-community
     +-ajax
     |   +-rss
     |   +-social
     +-culture

+-computer
     +-howto
     |   +-article
     |   +-diy
     |   +-hacks
     |   +-security
     |   +-tips
     |   +-tutorial
     |   |     +-tutorials
     |   +-webdesign
     |        +-ajax
     |        |   +-javascript
     |        +-blogs
     |        |   +-business
     |        |   +-community
     |        |   |     +-social
     |        |   +-news
     |        |   +-rss
     |        +-css
     |        |  +-html
     |        +-development
     |        |      +-code
     |        |      +-programming
     |        +-firefox
     |        +-inspiration
     |        |      +-art
     |        |        +-graphics
     |        +-library
     |        +-php
     |        +-tool
     |        |   +-download
     |        |   |     +-freeware
     |        |   |     +-mac
     |        |   |     |  +-apple
     |        |   |     |  +-osx
     |        |   |     +-opensource
     |        |   |     |      +-linux
     |        |   |     +-windows
     |        |   +-online
     |        |   |    +-media
     |        |   |       +-video
     |        |   +-productivity
     |        |   |       +-lifehacks
     |        |   +-utilities
     |        +-webdev
     +-interesting
          +-fun
          +-research
               +-education
               +-search

+-cool
   +-interesting
   |      +-fun
   |      |  +-funny
   |      |     +-humor
   |      +-media
   |      |   +-community
   |      |   |     +-social
   |      |   +-culture
   |      |   +-flash
   |      |   |   +-games
   |      |   +-music
   |      |   +-news
   |      |   |   +-rss
   |      |   +-video
   |      +-research
   |      |     +-business
   |      |     +-education
   |      |     |     +-science
   |      |     +-search
   |      +-writing
   +-software
        +-computer
        +-design
        |    +-art
        |    |  +-graphics
        |    |  +-images
        |    |      +-photography
        |    |            +-photo
        |    |               +-photos
        |    +-blog
        |    |   +-blogs
        |    +-inspiration
        |    +-resources
        |    |     +-howto
        |    |        +-article
        |    |        +-diy
        |    |        +-hacks
        |    |        +-tips
        |    |        +-tutorial
        |    |             +-tutorials
        |    +-webdesign
        |         +-webdev
        +-development
        |      +-ajax
        |      |   +-javascript
        |      +-code
        |      +-css
        |      |  +-html
        |      +-programming
        +-firefox
        +-free
        |   +-download
        |        +-freeware
        +-google
        +-internet
        |     +-online
        |     +-web2.0
        +-library
        +-opensource
        |      +-linux
        +-productivity
        |       +-lifehacks
        +-tech
        |   +-technology
        +-tool
        |   +-utilities
        +-windows
            +-mac

+-css
  +-html
     +-graphics
     +-javascript
```

```
     +-php

+-culture
    +-funny
       +-humor

+-design
    +-blog
    |   +-blogs
    |   +-diy
    |   +-interesting
    |   |       +-fun
    |   |       |  +-art
    |   |       |  |  +-graphics
    |   |       |  |  +-images
    |   |       |  |       +-photography
    |   |       |  |             +-photo
    |   |       |  |             +-photos
    |   |       |  +-funny
    |   |       +-inspiration
    |   |       +-media
    |   |          +-culture
    |   |          +-news
    |   |          |   +-community
    |   |          |         +-social
    |   |          +-video
    |   +-resources
    |   |      +-free
    |   |      |   +-download
    |   |      |   |     +-freeware
    |   |      |   +-online
    |   |      |   +-opensource
    |   |      |   +-tool
    |   |      |      +-utilities
    |   |      +-howto
    |   |      |   +-article
    |   |      |   +-hacks
    |   |      |   +-tips
    |   |      |   +-tutorial
    |   |      |         +-tutorials
    |   |      +-library
    |   |      +-php
    |   |        +-java
    |   +-rss
    |   +-technology
    |   |       +-computer
    |   |       +-firefox
    |   |       +-mac
    |   |       +-productivity
    |   |       |       +-lifehacks
    |   |       |       +-writing
    |   |       +-research
    |   |            +-business
    |   |            +-education
    |   |            +-search
    |   +-web2.0
    |       +-ajax
    |       |   +-javascript
    |       +-webdesign
    |            +-development
    |                  +-code
    |                  +-css
    |                  |  +-html
    |                  +-programming
    |                  +-webdev
    +-flash
       +-photoshop

+-development
      +-online
      |    +-blogs
      |    |   +-business
      |    |   +-inspiration
      |    |   |      +-graphics
      |    |   +-news
      |    |      +-community
      |    |           +-social
      |    +-fun
      |    +-hacks
      |    |   +-firefox
      |    |   +-productivity
      |    |   +-windows
      |    |      +-mac
      |    +-media
      |    +-research
      |    |     +-education
      |    +-search
      |    +-utilities
      |         +-download
      |              +-freeware
      |              +-opensource
      |                    +-linux
      +-programming
            +-code
            +-library
            |    +-java
            +-php
            +-tutorial
            |     +-article
            |     +-tutorials
            +-webdev
                +-ajax
                   +-css
                   |  +-html
                   +-javascript

+-download
     +-opensource
     |      +-freeware
     |      +-library
     |      +-windows
     |          +-linux
     |          +-mac
     |            +-osx
     +-search
        +-video

+-education
     +-culture
        +-science

+-free
   +-computer
   |     +-download
   |     |     +-freeware
   |     |     +-opensource
   |     |     |      +-linux
   |     |     +-windows
   |     |         +-mac
   |     |           +-osx
   |     +-firefox
   |     +-howto
   |     |   +-article
   |     |   +-diy
   |     |   +-hacks
   |     |   +-tips
   |     |   +-tutorial
   |     |         +-tutorials
   |     +-productivity
   |               +-lifehacks
   +-web2.0
       +-ajax
       |   +-javascript
       +-blogs
       |   +-news
       |   +-rss
       +-community
       |     +-social
       +-inspiration
       |      +-art
       |        +-graphics
       |             +-images
       +-interesting
```

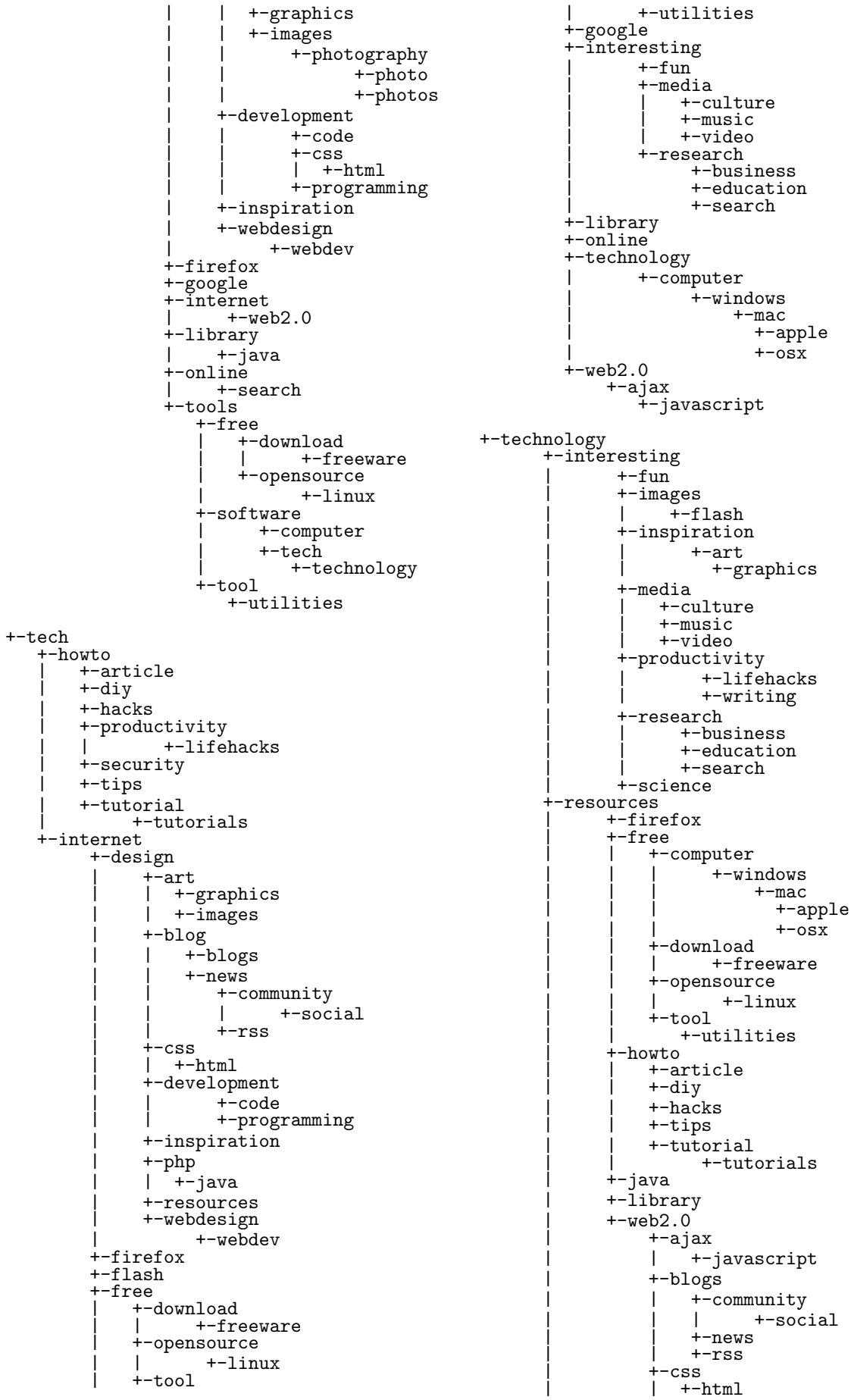
```
       |      +-fun
       |      +-research
       |           +-business
       |           +-education
       |           +-search
       +-library
       +-media
       |   +-flash
       |   +-music
       |   +-video
       +-online
       +-tool
       |   +-utilities
       +-webdesign
            +-css
            |  +-html
            +-development
            |      +-code
            |      +-programming
            +-webdev

+-freeware
     +-hacks
        +-productivity
        +-windows
            +-linux
            +-mac
              +-osx

+-fun
  +-blogs
  |   +-community
  |   |     +-social
  |   +-inspiration
  |   |      +-art
  |   |      |  +-graphics
  |   |      |       +-images
  |   |      +-photography
  |   |           +-photo
  |   |              +-photos
  |   +-media
  |   |   +-culture
  |   |   |    +-funny
  |   |   |       +-humor
  |   |   +-music
  |   |   +-video
  |   +-news
  |   +-research
  |   |     +-business
  |   |     +-education
  |   |     |     +-science
  |   |     +-search
  |   +-tutorials
  |        +-article
  |        +-diy
  |        +-hacks
  |           +-lifehacks
  +-utilities
       +-download
       |     +-freeware
       +-flash
          +-games

+-funny
   +-humor

+-graphics
     +-images
         +-photography
         |      +-photo
         |      +-photos
         +-photoshop

+-hacks
   +-productivity
         +-diy
         +-lifehacks

+-howto
   +-interesting
   |      +-fun
   |      +-media
   |      +-research
   |      |     +-education
   |      |     +-search
   |      +-tips
   |         +-article
   |         +-hacks
   |         |   +-diy
   |         +-tutorial
   |              +-tutorials
   +-webdesign
        +-ajax
        |   +-javascript
        +-blogs
        |   +-business
        |   |     +-productivity
        |   |           +-lifehacks
        |   +-inspiration
        |   |      +-art
        |   |        +-graphics
        |   +-news
        |      +-community
        +-css
        |  +-html
        +-development
        |      +-code
        |      +-programming
        +-library
        +-php
        +-tool
        |   +-download
        |   |     +-freeware
        |   |     +-opensource
        |   |           +-linux
        |   +-online
        |   +-utilities
        |   +-windows
        |       +-mac
        +-webdev

+-html
   +-graphics
   +-javascript
         +-php

+-images
    +-photography
          +-photo
          +-photos

+-imported
     +-reference
     |     +-library
     |     |    +-php
     |     |      +-java
     |     +-research
     |     |     +-business
     |     |     +-education
     |     |     +-science
     |     |     +-search
     |     +-resources
     |     +-toread
     |     |    +-howto
     |     |       +-article
     |     |       +-tips
     |     |       +-tutorial
     |     |            +-tutorials
     |     +-writing
     |         +-books
     |         +-language
     +-shopping
     |     +-games
     +-web
     |  +-cool
     |  |   +-flash
```


+-fun
+-funny
+-humor
+-interesting
+-media
+-community
+-social
+-culture
+-music
+-audio
+-news
+-video
+-design
+-art
+-graphics
+-images
+-photography
+-photo
+-photos
+-blog
+-blogs
+-rss
+-development
+-code
+-css
+-html
+-programming
+-inspiration
+-webdesign
+-webdev
+-google
+-internet
+-web2.0
+-ajax
+-javascript
+-tools
+-free
+-download
+-freeware
+-online
+-opensource
+-linux
+-hacks
+-diy
+-productivity
+-lifehacks
+-software
+-computer
+-tech
+-technology
+-tool
+-utilities
+-windows
+-firefox
+-mac
+-apple
+-osx
+-security
+-inspiration
+-article
+-code
+-ajax
+-css
+-html
+-research
+-business
+-productivity
+-education
+-media
+-art
+-graphics
+-images
+-photography
+-photo
+-photos
+-community
+-social
+-culture
+-news
+-search
+-download
+-funny
+-flash
+-interesting
+-fun
+-funny
+-humor
+-media
+-culture
+-video
+-webdesign
+-ajax
+-blogs
+-business
+-community
+-social
+-news
+-rss
+-css
+-html
+-development
+-code
+-programming
+-download
+-freeware
+-opensource
+-flash
+-inspiration
+-art
+-graphics
+-images
+-research
+-education
+-science
+-search
+-writing
+-tips
+-article
+-diy
+-hacks
+-productivity
+-lifehacks
+-tutorial
+-tutorials
+-tool
+-online
+-utilities
+-webdev
+-internet
+-design
+-blog
+-blogs
+-development
+-code
+-css
+-html
+-programming
+-webdev
+-flash
+-inspiration
+-php
+-resources
+-howto
+-article
+-hacks
+-tips
+-tutorial
+-tutorials
+-technology
+-computer

```
|      +-diy
|      +-firefox
|      +-free
|      |   +-download
|      |   |     +-freeware
|      |   +-library
|      |   +-online
|      |   +-opensource
|      |   |      +-linux
|      |   +-tool
|      |   |   +-utilities
|      |   +-windows
|      |       +-mac
|      +-google
|      +-interesting
|            +-fun
|            |  +-funny
|            +-media
|            |   +-art
|            |   |  +-graphics
|            |   |  +-images
|            |   +-community
|            |   |     +-social
|            |   +-culture
|            |   +-music
|            |   +-news
|            |   |   +-rss
|            |   +-video
|            +-productivity
|            |       +-lifehacks
|            +-research
|            |     +-business
|            |     +-education
|            |     +-search
|            +-writing
|      +-web2.0
|      |    +-ajax
|      |       +-javascript
|      +-webdesign

+-javascript
      +-php
        +-java

+-lifehacks
     +-diy
     +-gtd

+-linux
   +-mac

+-mac
  +-osx
    +-apple

+-media
   +-art
   |  +-graphics
   |       +-images
   +-business
   |     +-education
   |     +-news
   |     |   +-culture
   |     |   +-rss
   |     +-search
   |     |    +-community
   |     |         +-social
   |     +-utilities
   |          +-download
   |               +-freeware
   +-video
      +-music
         +-audio

+-music
   +-audio

+-news
   +-art
     +-community
     |     +-rss
     |     +-social
     +-culture
     +-education

+-online
    +-fun
    |  +-blogs
    |  |   +-community
    |  |   |     +-social
    |  |   +-firefox
    |  |   +-google
    |  |   +-inspiration
    |  |   |      +-art
    |  |   |        +-graphics
    |  |   |             +-images
    |  |   +-media
    |  |   |   +-culture
    |  |   |   +-music
    |  |   |   +-video
    |  |   +-news
    |  |   +-research
    |  |   |     +-business
    |  |   |     +-education
    |  |   |     +-library
    |  |   |     +-search
    |  |   +-rss
    |  +-flash
    +-tutorial
        +-article
        +-download
        |     +-freeware
        |     +-opensource
        +-hacks
        +-tutorials
        +-webdev
            +-ajax
            +-code
            +-css
            +-html
            +-utilities
                 +-productivity
                        +-lifehacks

+-opensource
      +-community
      |     +-ajax
      |        +-html
      |        +-library
      +-freeware
           +-mac
           +-windows
               +-linux

+-osx
  +-apple

+-photo
   +-photos

+-photography
      +-photo
         +-photos

+-productivity
       +-lifehacks
            +-gtd

+-programming
      +-online
      |    +-blogs
      |    |   +-business
      |    |   +-inspiration
      |    |   |      +-graphics
      |    |   +-news
      |    +-community
```

```
|    +-fun
|    +-research
|    |     +-education
|    +-search
|    +-utilities
|         +-download
|         |     +-freeware
|         |     +-opensource
|         |           +-linux
|         +-windows
|            +-mac
 +-tutorial
      +-article
      +-hacks
      |   +-firefox
      |   +-productivity
      +-library
      |    +-java
      +-tutorials
      +-webdev
          +-ajax
          |   +-css
          |   |  +-html
          |   +-javascript
          +-code
          +-php

+-reference
     +-toread
     |    +-blog
     |    |   +-blogs
     |    |   +-rss
     |    +-hacks
     |    |   +-diy
     |    |   +-productivity
     |    |   |       +-gtd
     |    |   |       +-lifehacks
     |    |   +-windows
     |    |       +-mac
     |    |       +-security
     |    +-howto
     |       +-article
     |       +-tips
     |       +-tutorial
     |            +-tutorials
     +-web
       +-ajax
       |   +-java
       |   +-javascript
       |   +-php
       +-cool
       |   +-flash
       |   +-fun
       |   +-funny
       |   +-interesting
       +-design
       |    +-art
       |    |  +-graphics
       |    |  +-images
       |    |  +-photography
       |    |        +-photo
       |    |        +-photos
       |    +-development
       |    |      +-code
       |    |      +-css
       |    |      |  +-html
       |    |      +-programming
       |    +-inspiration
       |    +-webdesign
       |         +-webdev
       +-firefox
       +-internet
       |     +-web2.0
       +-research
       |     +-books
       |     +-business
       |     +-education
       |     |     +-science
       |     +-google
       |     +-language
       |     +-library
       |     +-media
       |     |   +-community
       |     |   |     +-social
       |     |   +-culture
       |     |   +-music
       |     |   +-news
       |     |   +-video
       |     +-search
       |     +-writing
      +-tools
         +-free
         |   +-download
         |   |     +-freeware
         |   +-online
         +-resources
         +-software
         |     +-computer
         |     +-opensource
         |     |      +-linux
         |     +-tech
         |        +-technology
         +-tool
            +-utilities

+-research
     +-media
     |   +-art
     |   +-business
     |   +-community
     |   |     +-social
     |   +-culture
     |   +-education
     |   |     +-science
     |   |     +-writing
     |   +-news
     +-utilities
          +-ajax
          +-opensource
          |      +-library
          +-search

+-resources
     +-free
     |   +-computer
     |   |     +-mac
     |   |     +-windows
     |   +-download
     |   |     +-freeware
     |   +-firefox
     |   +-flash
     |   +-library
     |   +-online
     |   +-opensource
     |   |      +-linux
     |   +-tool
     |   |   +-utilities
     |   +-web2.0
     |       +-ajax
     |       |   +-javascript
     |       +-blogs
     |       |   +-inspiration
     |       |   |      +-art
     |       |   |        +-graphics
     |       |   |        +-images
     |       |   +-news
     |       |   |   +-community
     |       |   |        +-social
     |       |   +-rss
     |       +-php
     |       +-webdesign
     |            +-css
     |            |  +-html
```

```
|            +-development
|            |      +-code
|            |      +-programming
|            +-webdev
+-howto
   +-article
   +-diy
   +-hacks
   +-interesting
   |      +-fun
   |      +-media
   |      |   +-culture
   |      |   +-photography
   |      |   +-video
   |      +-productivity
   |      |       +-lifehacks
   |      +-research
   |      |     +-business
   |      |     +-education
   |      |     +-search
   |      +-writing
   |          +-books
   +-tips
   +-tutorial
        +-tutorials

+-search
    +-news
       +-community
       |     +-social
       +-education
       |     +-library
       +-google

+-social
    +-culture

+-software
     +-free
     |   +-download
     |   |     +-freeware
     |   +-flash
     |   +-library
     |   +-opensource
     |   |      +-linux
     |   +-productivity
     |   |       +-lifehacks
     |   +-tool
     |      +-utilities
     +-tech
        +-computer
        |     +-windows
        |         +-mac
        |           +-apple
        |           +-osx
        +-howto
        |   +-article
        |   +-diy
        |   +-education
        |   +-hacks
        |   +-tips
        |   +-tutorial
        |        +-tutorials
        +-internet
        |     +-design
        |     |    +-art
        |     |    |  +-graphics
        |     |    |  +-images
        |     |    +-blog
        |     |    |   +-blogs
        |     |    |   +-news
        |     |    |      +-community
        |     |    |      |     +-social
        |     |    |      +-rss
        |     |    +-css
        |     |    |  +-html
        |     |    +-development
        |     |    |      +-code
        |     |    |      +-programming
        |     |    +-inspiration
        |     |    +-php
        |     |    |  +-java
        |     |    +-resources
        |     |    +-webdesign
        |     |         +-webdev
        |     +-firefox
        |     +-google
        |     +-interesting
        |     |      +-fun
        |     |      +-research
        |     |           +-business
        |     |           +-search
        |     +-media
        |     |   +-video
        |     +-online
        |     +-web2.0
        |         +-ajax
        |            +-javascript
        +-security
        +-technology

+-system:unfiled
        +-imported
            +-games
            +-reference
            |     +-hacks
            |     |   +-diy
            |     |   +-windows
            |     |       +-mac
            |     |       |  +-apple
            |     |       |  +-osx
            |     |       +-security
            |     +-research
            |     |     +-business
            |     |     +-education
            |     |     +-science
            |     +-resources
            |     +-toread
            |     |    +-blog
            |     |    |   +-blogs
            |     |    |   +-rss
            |     |    +-howto
            |     |    |   +-article
            |     |    |   +-tips
            |     |    |   +-tutorial
            |     |    |        +-tutorials
            |     |    +-productivity
            |     |           +-gtd
            |     |           +-lifehacks
            |     +-writing
            |         +-books
            |         +-language
            +-shopping
            +-web
              +-ajax
              |   +-javascript
              |   +-php
              +-cool
              |   +-flash
              |   +-fun
              |   +-funny
              |   |   +-humor
              |   +-interesting
              |   +-media
              |      +-community
              |      |     +-social
              |      +-culture
              |      +-music
              |      |   +-audio
              |      +-news
              |      +-video
              +-design
              |    +-art
```


| +-graphics
| +-images
| +-photography
| +-photo
| +-photos
+-development
+-code
+-css
| +-html
+-programming
+-inspiration
+-webdesign
+-webdev
+-firefox
+-google
+-internet
+-web2.0
+-library
+-java
+-online
+-search
+-tools
+-free
+-download
+-freeware
+-opensource
+-linux
+-software
+-computer
+-tech
+-technology
+-tool
+-utilities
+-tech
+-howto
+-article
+-diy
+-hacks
+-productivity
+-lifehacks
+-security
+-tips
+-tutorial
+-tutorials
+-internet
+-design
+-art
+-graphics
+-images
+-blog
+-blogs
+-news
+-community
+-social
+-rss
+-css
+-html
+-development
+-code
+-programming
+-inspiration
+-php
+-java
+-resources
+-webdesign
+-webdev
+-firefox
+-flash
+-free
+-download
+-freeware
+-opensource
+-linux
+-tool
+-utilities
+-google
+-interesting
+-fun
+-media
+-culture
+-music
+-video
+-research
+-business
+-education
+-search
+-library
+-online
+-technology
+-computer
+-windows
+-mac
+-apple
+-osx
+-web2.0
+-ajax
+-javascript
+-technology
+-interesting
+-fun
+-images
+-flash
+-inspiration
+-art
+-graphics
+-media
+-culture
+-music
+-video
+-productivity
+-lifehacks
+-writing
+-research
+-business
+-education
+-search
+-science
+-resources
+-firefox
+-free
+-computer
+-windows
+-mac
+-apple
+-osx
+-download
+-freeware
+-opensource
+-linux
+-tool
+-utilities
+-howto
+-article
+-diy
+-hacks
+-tips
+-tutorial
+-tutorials
+-java
+-library
+-web2.0
+-ajax
+-javascript
+-blogs
+-community
+-social
+-news
+-rss
+-css
+-html

```
      |         |  +-php
      |         +-google
      |         +-online
      |         +-webdesign
      |              +-development
      |              |      +-code
      |              |      +-programming
      |              +-webdev
      +-security

+-tips
   +-development
   |      +-ajax
   |      |   +-javascript
   |      +-blogs
   |      |   +-business
   |      |   |     +-productivity
   |      |   |            +-gtd
   |      |   |            +-lifehacks
   |      |   +-inspiration
   |      |   |      +-art
   |      |   |        +-graphics
   |      |   +-news
   |      |      +-community
   |      +-code
   |      +-css
   |      |  +-html
   |      +-opensource
   |      |      +-linux
   |      |      +-mac
   |      +-programming
   |      +-tutorial
   |      |     +-article
   |      |     +-tutorials
   |      +-webdev
   +-hacks
   |   +-diy
   +-online
       +-education
       +-fun
       +-media
       +-research
       +-search
       +-utilities
            +-download
            |     +-freeware
            +-windows

+-tool
   +-online
   |    +-blogs
   |    |   +-community
   |    |   |     +-social
   |    |   +-inspiration
   |    |   |      +-art
   |    |   |        +-graphics
   |    |   |             +-images
   |    |   +-news
   |    |   +-rss
   |    +-fun
   |    +-google
   |    +-media
   |    +-research
   |    |     +-business
   |    |     +-education
   |    |     +-search
   |    +-utilities
   |         +-download
   |         |     +-freeware
   |         |     +-mac
   |         |     |  +-osx
   |         |     +-opensource
   |         |     |      +-linux
   |         |     +-windows
   |         +-firefox
   |         +-productivity
   |               +-lifehacks
   +-tips
      +-article
      +-development
      |      +-ajax
      |      |   +-javascript
      |      +-code
      |      +-css
      |      |  +-html
      |      +-library
      |      +-php
      |      +-programming
      |      +-webdev
      +-diy
      +-hacks
      +-tutorial
          +-tutorials

+-tools
   +-toread
       +-cool
       |   +-design
       |   |    +-art
       |   |    |  +-graphics
       |   |    |  |     +-images
       |   |    |  +-photography
       |   |    |         +-photo
       |   |    |            +-photos
       |   |    +-blog
       |   |    |   +-blogs
       |   |    +-development
       |   |    |      +-code
       |   |    |      +-css
       |   |    |      |  +-html
       |   |    |      +-programming
       |   |    +-inspiration
       |   |    +-php
       |   |    |  +-java
       |   |    +-resources
       |   |    +-webdesign
       |   |         +-webdev
       |   +-flash
       |   +-interesting
       |   |      +-fun
       |   +-internet
       |   |     +-online
       |   |     +-web2.0
       |   |         +-ajax
       |   |            +-javascript
       |   +-media
       |   |   +-culture
       |   |   +-music
       |   |   +-video
       |   +-research
       |        +-business
       |        +-education
       |        +-google
       |        +-library
       |        +-news
       |        |   +-community
       |        |   |     +-social
       |        |   +-rss
       |        +-search
       |        +-writing
       +-hacks
       |   +-diy
       +-howto
       |   +-article
       |   +-tips
       |   +-tutorial
       |         +-tutorials
       +-productivity
       |       +-gtd
       |       +-lifehacks
       +-software
```

```
            +-firefox
            +-free
            |   +-download
            |   |     +-freeware
            |   +-opensource
            |   |      +-linux
            |   +-tool
            |      +-utilities
            +-security
            +-tech
            |   +-computer
            |   +-technology
            +-windows
                +-mac
                  +-apple
                  +-osx

+-toread
    +-cool
    |   +-design
    |   |    +-ajax
    |   |    |   +-css
    |   |    |   |  +-html
    |   |    |   +-javascript
    |   |    +-art
    |   |    |  +-graphics
    |   |    |  +-images
    |   |    +-blog
    |   |    |   +-blogs
    |   |    +-development
    |   |    |      +-code
    |   |    |      +-programming
    |   |    +-inspiration
    |   |    +-php
    |   |    |  +-java
    |   |    +-resources
    |   |    +-webdesign
    |   |         +-webdev
    |   +-fun
    |   |  +-funny
    |   +-interesting
    |   +-research
    |        +-business
    |        |     +-productivity
    |        |            +-gtd
    |        |            +-lifehacks
    |        +-education
    |        |     +-science
    |        +-google
    |        +-library
    |        +-media
    |        |   +-community
    |        |   |     +-social
    |        |   +-culture
    |        |   +-news
    |        |   +-rss
    |        |   +-video
    |        +-search
    |        +-writing
    |           +-books
    +-software
         +-computer
         +-firefox
         +-free
         |   +-download
         |   +-freeware
         |   +-online
         |   +-opensource
         |   |      +-linux
         |   +-tool
         |      +-utilities
         +-howto
         |   +-article
         |   +-diy
         |   +-hacks
         |   +-tips
         |   +-tutorial
         |       +-tutorials
         +-internet
         |     +-web2.0
         +-tech
         |   +-technology
         +-windows
             +-mac

+-tutorial
     +-fun
     |  +-diy
     |  +-research
     |  |     +-education
     |  +-utilities
     |       +-download
     |       |     +-freeware
     |       |     +-opensource
     |       |           +-linux
     |       +-hacks
     |          +-productivity
     |                +-lifehacks
     +-webdev
         +-ajax
         |   +-javascript
         +-blogs
         |   +-article
         |   +-business
         |   +-inspiration
         |         +-art
         |           +-graphics
         +-code
         +-css
         |  +-html
         +-library
         +-php
         +-tutorials

+-tutorials
     +-article
     |    +-code
     |       +-ajax
     |       |   +-css
     |       |   |  +-html
     |       |   +-javascript
     |       +-opensource
     |             +-linux
     +-inspiration
     |      +-art
     |      |  +-graphics
     |      +-research
     |           +-education
     +-utilities
          +-hacks
             +-diy
             +-productivity
                    +-lifehacks

+-utilities
     +-code
     |   +-ajax
     |   +-html
     +-download
     |     +-freeware
     |     +-mac
     |     +-opensource
     |     |      +-linux
     |     +-windows
     +-hacks
     |   +-productivity
     |         +-lifehacks
     +-search
         +-firefox

+-video
   +-music

+-web
```

```
+-tools
|   +-free
|   |   +-download
|   |   |     +-freeware
|   |   +-online
|   |   +-opensource
|   |         +-linux
|   +-hacks
|   |   +-diy
|   |   +-productivity
|   |          +-lifehacks
|   +-internet
|   |     +-web2.0
|   |         +-ajax
|   |            +-javascript
|   +-resources
|   +-software
|   |     +-tech
|   |        +-computer
|   |        +-technology
|   +-tool
|   |   +-utilities
|   +-windows
|       +-firefox
|       +-mac
+-toread
    +-cool
    |   +-design
    |   |    +-blog
    |   |    |   +-blogs
    |   |    |   +-rss
    |   |    +-development
    |   |    |      +-code
    |   |    |      +-css
    |   |    |      |  +-html
    |   |    |      +-programming
    |   |    |      +-webdev
    |   |    +-images
    |   |    |    +-photography
    |   |    |         +-photo
    |   |    |         +-photos
    |   |    +-inspiration
    |   |    |      +-art
    |   |    |        +-graphics
    |   |    +-php
    |   |    |  +-java
    |   |    +-webdesign
    |   +-flash
    |   +-fun
    |   |  +-funny
    |   |     +-humor
    |   +-interesting
    |   +-media
    |      +-community
    |      |     +-social
    |      +-culture
    |      +-music
    |      +-news
    |      +-video
    +-howto
    |   +-article
    |   +-tips
    |   +-tutorial
    |        +-tutorials
    +-research
         +-business
         +-education
         +-google
         +-library
         +-search
         +-writing
             +-books

+-web2.0
   +-computer
   |     +-firefox
   |     +-howto
   |     |   +-article
   |     |   +-hacks
   |     |   +-php
   |     |   +-tips
   |     |   +-tutorial
   |     |   |     +-tutorials
   |     |   +-webdesign
   |     |        +-ajax
   |     |        |   +-javascript
   |     |        +-development
   |     |        |      +-code
   |     |        |      +-css
   |     |        |      |  +-html
   |     |        |      +-programming
   |     |        |      +-webdev
   |     |        +-inspiration
   |     |               +-art
   |     |                 +-graphics
   |     |                 +-images
   |     +-interesting
   |     |      +-blogs
   |     |      |   +-news
   |     |      |   +-rss
   |     |      +-fun
   |     |      |  +-media
   |     |      |     +-video
   |     |      +-google
   |     |      +-research
   |     |           +-business
   |     |           +-community
   |     |           |     +-culture
   |     |           |     +-social
   |     |           +-education
   |     |           +-search
   |     +-library
   |     +-tool
   |        +-download
   |        |     +-freeware
   |        |     +-mac
   |        |     +-opensource
   |        +-online
   |        +-productivity
   |        |       +-lifehacks
   |        +-utilities
   +-flash

+-webdesign
    +-tips
    |   +-article
    |   +-development
    |   |      +-ajax
    |   |      |   +-javascript
    |   |      +-css
    |   |      |  +-html
    |   |      +-php
    |   |      +-programming
    |   |      |      +-code
    |   |      +-webdev
    |   +-hacks
    |   +-tutorial
    |        +-tutorials
    +-tool
       +-download
       |     +-freeware
       |     +-mac
       |     +-opensource
       +-firefox
       +-fun
       |  +-blogs
       |  |   +-news
       |  |   |   +-community
       |  |   |        +-social
       |  |   +-research
       |  |   |     +-business
       |  |   |     +-education
```

```
        |  |   |     +-search
        |  |   +-rss
        |  +-flash
        |  +-inspiration
        |  |      +-art
        |  |        +-graphics
        |  |        +-images
        |  +-media
        |  +-photography
        |        +-photo
        |        +-photos
        +-library
        |    +-java
        +-online
        +-productivity
        +-utilities

+-webdev
    +-fun
    |  +-blogs
    |  |   +-inspiration
    |  |   |      +-art
    |  |   |      |  +-graphics
    |  |   |      +-css
    |  |   |        +-html
    |  |   +-news
    |  |   +-research
    |  |        +-business
    |  |        +-community
    |  |        +-search
    |  +-utilities
    |       +-download
    |       |     +-freeware
    |       |     +-opensource
    |       +-productivity
    +-linux
    +-tutorials
         +-ajax
         |   +-java
         |   +-javascript
         |   +-library
         |   +-php
         +-article
         +-code
         +-hacks
            +-firefox

+-windows
    +-linux
       +-mac
```

参考文献

1) Netcraft. http://news.netcraft.com.

2) Netcraft. December 2007 web server survey. http://news.netcraft.com/archives/2007/12/29/december_2007_web%server_survey.html.

3) Kosala and Blockeel. Web mining research: A survey. *SIGKDD: SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining, ACM*, 2, 2000.

4) G.Smith. Atomiq: Folksonomy: social classification, August 2004. http://atomiq.org/archives/2004/08/folksonomy_social_classification.html.

5) Yahoo. http://www.yahoo.co.jp/index.html.

6) 丹羽 智史, 土肥 拓生, and 本位田 真一. Folksonomy の 3 部グラフ構造を利用したタグクラスタリング. **人工知能学会 セマンティックウェブとオントロジー研究会**, November 2006.

7) del.icio.us. http://del.icio.us/.

8) Digg. http://digg.com/.

9) はてなブックマーク. http://b.hatena.ne.jp/.

10) flickr. http://www.flickr.com/.

11) YouTube. http://www.youtube.com/.

12) citeulike. http://www.citeulike.org/.

13) Connotea. http://www.connotea.org/.

14) Andreas Hotho, Robert Jaschke, Christoph Schmitz, and Gerd Stumme. Trend detection in folksonomies. In Yannis S. Avrithis, Yiannis Kompatsiaris, Steffen Staab, and Noel E. O'Connor, editors, *Proc. First International Conference on Semantics And Digital Media Technology (SAMT)*, volume 4306 of *LNCS*, pages 56–70, Heidelberg, dec 2006. Springer.

15) Micah Dubinko, Ravi Kumar, Joseph Magnani, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. Visualizing tags over time. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 193–202, New York, NY, USA, 2006. ACM Press.

16) S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, 1998.

17) Shenghua Bao, Guirong Xue, Xiaoyuan Wu, Yong Yu, Ben Fei, and Zhong Su. Optimizing web search using social annotations. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 501–510, New York, NY, USA, 2007. ACM Press.

18) T. Berners-Lee, J. Hendler, and O. Lassila. The semantic Web. *Scientific American*, 284(5):28–37, 2001.

19) P. Schmitz. Inducing ontology from flickr tags. In *Proc. of the Collaborative Web Tagging Workshop (WWW '06)*, May 2006.

20) L. Specia and E. Motta. Integrating Folksonomies with the Semantic Web. *Proc. of ESWC*, 7, 2007.

21) Simone Braun, Andreas Schmidt, Andreas Walter, Gabor Nagypal, and Valentin Zacharias. Ontology maturing: a collaborative web 2.0 approach to ontology engineering. In Natasha Noy, Harith Alani, Gerd Stumme, Peter Mika, York Sure, and Denny Vrandecic, editors, *Proceedings of the Workshop on Social and Collaborative Construction of Structured Knowledge (CKC 2007) at the 16th International World Wide Web Conference (WWW2007) Banff, Canada, May 8, 2007*, volume 273 of *CEUR Workshop Proceedings*, 2007.

22) Paul Heymann and Hector Garcia-Molina. Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical Report 2006-10, Stanford University, April 2006.

23) Grigory Begelman, Philipp Keller, and Frank Smadja. Automated tag clustering: Improving search and exploration in the tag space. 2006.

24) Miranda Grahl, Andreas Hotho, and Gerd Stumme. Conceptual clustering of social bookmarking sites. In *Proc. I-Know 2007 Conference (to appear)*, Graz, Austria, September 2007.

25) Open Directory Project. http://www.dmoz.org/.

26) D. Haussler. Convolution kernels on discrete structures, 1999.

27) H. Kashima and T. Koyanagi. Kernels for semi-structured data. *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 291–298, 2002.

28) 久保山 哲二, 申 吉浩, 鹿島 久嗣, and 平田 耕一. 共通構造の数え上げによる半構造データカーネルの設計.

29) P. Chan. A non-invasive learning approach to building web user profiles. *Workshop on Web usage analysis and user profiling, Fifth International Conference on Knowledge Discovery and Data Mining, San Diego.*, August 1999.

30) C. J. Van Rijsbergen. *Information Retrieval, 2nd edition.* Dept. of Computer Science, University of Glasgow, 1979.

31) R. Rosenfeld. A maximum entropy approach to adaptive statistical language modeling, 1996.

32) C. Fellbaum. *Wordnet: An Electronic Lexical Database.* MIT Press, 1998.

33) 鹿島 久嗣, 坂本 比呂志, and 小柳 光生. 木構造データに対するカーネル関数の設計と解析. **人工知能学会論文誌**, 21(1):113–121, 2006.

34) C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

35) P. Cimiano, S. Handschuh, and S. Staab. Towards the self-annotating web, 2004.

36) Philipp Cimiano, Günter Ladwig, and Steffen Staab. Gimme' the context: context-driven automatic semantic annotation with c-pankow. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 332–341, New York, NY, USA, 2005. ACM Press.

37) A.A. Patil, S.A. Oundhakar, A.P. Sheth, and K. Verma. Meteor-s web service annotation framework. *Proceedings of the 13th conference on World Wide Web*, pages 553–562, 2004.

38) Andreas Heß and Nicholas Kushmerick. Automatically attaching semantic metadata to web services. *Workshop on Information Integration on the Web (IIWeb)*, 2003.

39) A. Heß and N. Kushmerick. Machine Learning for Annotating Semantic Web Services. *AAAI Spring Symposium on Semantic Web Services*, 2004.

40) Paul, Stefania Costache, Wolfgang Nejdl, and Siegfried Handschuh. P-tag: large scale automatic generation of personalized annotation tags for the web. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 845–854, New York, NY, USA, 2007. ACM Press.

41) Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

42) Thorsten Joachims. Text categorization with support vector machines: learning with many relevant features. In Claire Nédellec and Céline Rouveirol, editors, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, number 1398, pages 137–142, Chemnitz, DE, 1998. Springer Verlag, Heidelberg, DE.

43) M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

44) Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5):513–523, 1988.

45) Chih-Jen Lin, Ruby C. Weng, and S. Sathiya Keerthi. Trust region Newton method for large-scale logistic regression. In *Proceedings of the 24th International Conference on Machine Learning (ICML)*, 2007.

46) Chih C. Chang and Chih J. Lin. *LIBSVM: a library for support vector machines*, 2001.

発表文献

1) Ken Ogino and Hiroyuki Sato. Extracting tag hierarchy from folksonomy. In *SWWS*, pages 120–125, 2007.

謝辞

本研究を進めるにあたり，多くの方にお世話になりました．

指導教員である佐藤周行準教授には，研究環境を与えて頂いたことをはじめ，論文の作成やプレゼンテーションの方法など，多くのご指導，ご教授を頂きました．

金田研究室の金田康正教授には，他研究室である私に快く交流を開いて頂きました．

金田研究室の黒田久泰助教授には，研究の内外を問わず，多くの知識や技術をご教授頂きました。

金田研究室秘書の亀田文美代氏には私生活における健康面での相談にのって頂きました．おかげ様で修士課程の二年間を大事なく過ごすことができました．

吉田仁氏には論文の執筆において多くの助言を賜った他，非常に綿密な論文の添削をして頂きました．

他の金田・佐藤研の皆様にも公私にわたり多くの助言を頂きました．

ここに，心よりの感謝の意を表します．

平成 20 年 1 月 29 日