TÉCNICAS DE ROTULAGEM DE RECURSOS EM SISTEMAS DE RI

MÁRCIA SAMPAIO LIMA

TÉCNICAS DE ROTULAGEM DE RECURSOS EM SISTEMAS DE RI

Projeto de tese apresentado ao Programa de Pós-Graduação em Informática do Instituto de Computação da Universidade Federal do Amazonas como requisito parcial para a obtenção do grau de Doutor em Informática.

Orientador: Edleno Silva de Moura

Manaus

Novembro de 2011

Sumário

Capítulo 1

Introdução

A *World Wide Web* (*Web*) é considerada um repositório universal do conhecimento e da cultura humana que viabiliza o compartilhamento de ideias e de informações numa proporção jamais vista. Seu grande sucesso é decorrente da facilidade com que usuários publicam, acessam e alteram recursos *Web*(fotos, vídeos, livros, páginas), pois nenhum conhecimento técnico profundo é exigido destes [2].

Atualmente, os sistemas de rotulagem colaborativa (*folksonomies*) são o destaque da *Web 2.0* [20]. Esses sistemas permitem que usuários armazenem, compartilhem e rotulem diversos recursos. Eles são considerados fonte de informação valiosa já que agrupam interesses, preferências e contribuições de milhares de usuários [20].

Neste trabalho expomos um estudo das recentes pesquisas feitas na área de Recuperação de Informação (RI) sobre rotulagem de recursos da *Web* bem como apresentamos dois métodos de recomendação automática de *tags* propostos e avaliados por nós.

Rotular consiste em associar a um recurso pequenas palavras que o descrevam de forma simples e objetiva. Estas palavras são chamadas de rótulos ou *tags* e são frequentemente usadas para auxiliar a organização e a navegação entre os recursos. Existem dois tipos de rotulagem: (1) a colaborativa, onde um recurso pode ser rotulado por diversos usuários, o que ocorre em sites como o *Del.ici.ous*[1] onde diversos usuários podem rotular *sites*; e (2) a restrita, onde apenas um usuário aplica rótulos a um recurso, o que ocorre no *Flickr*[2] onde as fotos somente podem ser rotuladas por seus donos.

Porém, apesar das vantagens proporcionadas pelos sistemas de rotulagem colaborativa, Marek *et al*, em [14], afirma que rotular (*tagging*) é uma tarefa difícil para

[1] http://delicious.com/
[2] http://www.flickr.com/

os usuários, pois estes devem listar um conjunto de *tags* para cada recurso a ser rotulado. Assim, surgem os sistemas de recomendação automática de *tags* que facilitam a rotulagem sugerindo *tags* susceptíveis a escolha do usuário.

Um de nossos objetivos durante o doutorado é propor e avaliar métodos de recomendação automática de *tags*. Dois métodos já propostos por nós, combinam diversas fontes de evidências textuais extraídas de uma página *Web* p, com o objetivo de gerar um conjunto de *tags* que represente p. A combinação das evidências utilizadas é feita automaticamente utilizando-se algoritmos genéticos (AG). Ambos os métodos serão explicados nos próximos capítulos desta proposta.

Outro objetivo para o doutorado é avaliar o impacto da utilização de *tags* em sistemas de recuperação de informação, mais especificamente explorando seu uso como nova fonte de evidência de relevância em problemas de busca e classificação, assim como *interfaces* alternativas em máquinas de busca. Observamos que vários são os métodos existentes para recomendar *tags*, porém queremos avaliar a viabilidade do uso tais *tags* e analisar o impacto que estas podem causar em sistemas de busca e classificação.

1.1 Trabalhos Relacionados

As *tags*, assim como as nuvens de *tags*, se tornaram um recurso comum em *sites* e páginas da *Web*. Com tanta popularidade, ambas são alvo frequente de estudo [19] [20] [12] [14].

Em http://www.cloudlet.com/, é possível obter e instalar um complemento (*add-on*) para o *Firefox* [3] que insere nuvens de *tag* sensíveis ao contexto na *interface* dos *sites* *Google* [4], *Yahoo* [5] e *Twitter* [6] para os usuários navegarem de forma mais eficiente através dos resultados da busca. A justificativa para se instalar o complemento é que não há mais necessidade de percorrer toda a lista de resultados da busca para compreender o assunto procurado, pois o complemento mostra as palavras-chaves relevantes aos resultados de busca, o que facilita a pesquisa e navegação do usuário. Já Kuo *et al* [12] criaram o *PubCloud* que gera uma nuvem de *tags* a partir de palavras extraídas dos *abstracts* de documentos biomédicos retornados pela de buscas feitas na *PubMed* [7] .

Uma forma de se anotar (*tagging*) um objeto é extrair destes seus termos mais relevantes [19]. Rafiei e Mendelzon [16] propõem um método de se determinar automaticamente os principais termos de uma página *Web* baseada nos valores de relevância

[3] http://www.mozilla.com/
[4] http://www.google.com/
[5] http://www.yahoo.com/
[6] http://twitter.com/
[7] http://www.ncbi.nlm.nih.gov/pubmed/

destes termos para aquela página. Para isso, foram generalizadas duas técnicas bastante utilizadas de cálculo de *ranking* de páginas *Web*, que são: (1) *PageRank* [7]; e (2) *Hubs and Authorits* [11], gerando dois algoritmos que determinam os termos relevantes de páginas *Web*: (1) algoritmo para computação dos termos relevantes de uma página com um nível de propagação de influência, generalização do *PageRank* [7]; e (2) algoritmo para computação dos termos relevantes de uma página com dois níveis de propagação de influência, generalização do *Hubs and Authorits* [11]. No modelo de propagação de influência em um nível, a reputação de um termo t em uma página p é expressa pela seguinte fórmula:

$$R^n(p,t) = \begin{cases} \frac{d}{N_t} + (1-d)\sum_{q \to p} \frac{R^{n-1}(q,t)}{O(q)} & \text{se t aparece em p ;} \\ (1-d)\sum_{q \to p} \frac{R^{n-1}(q,t)}{O(q)} & \text{caso contrário.} \end{cases}$$

onde d é a probabilidade de um surfista randômico, procurando por um termo t nas páginas da *Web*, pular para uma página escolhida randomicamente entre as que contêm o termo *t*, (1-d) a probabilidade de ele seguir um *link* de saída da página p, N_t é o número de páginas na *Web* que contém o termo *t*, q $\to$ p representa um *link* da página q para a página p e *O(q)* é o número de *links* de saída de p. Já no modelo de propagação de influência em dois níveis, a reputação de um termo t em uma página p é expressa utilizando duas métricas, a de *Hub* e de Autoridade, da seguinte forma:

$$H^n(p,t) = \begin{cases} \frac{d}{2N_t} + (1-d)\sum\limits_{q \to p} \frac{A^{n-1}(q,t)}{I(q)} & \text{se t aparece em p ;} \\ (1-d)\sum_{q \to p} \frac{A^{n-1}(q,t)}{I(q)} & \text{caso contrário.} \end{cases}$$

onde *I(q)* representa o número de *links* de entrada da página *q*. Com esses algoritmos Rafiei e Mendelzon tentam inferir os melhores termos descritores de uma página *p*. O objetivo de ambos, na descoberta dos principais termos descritores de *p*, é utilização desta informação na validação de páginas comerciais e pessoais, informando como estas páginas são conhecidas na *Web*. Porém, estes termos podem ser usados como *tags* para *p*.

Os sistemas de recomendação de *tags* podem ser divididos em recomendadores baseados em grafo (*graph-based recommender*) e baseados em conteúdo (*content-based recommenders*) [14]. Em nosso estudo priorizamos os recomendadores baseados em conteúdo, pois já propusemos um método que combina diversas fontes de evidências textuais extraídas de uma página *Web* com o objetivo de gerar um conjunto de *tags* que represente esta página. Lipczak e Milios [14], afirmam que a tarefa de anotar recursos (*tagging*) é difícil para os usuários, pois estes devem listar um conjunto de *tags* para

cada recurso a ser rotulado. Com o objetivo de amenizar este problema eles propõem um sistema híbrido de recomendação automática de *tags* que combina cinco diferentes listas de *tags* provenientes de cinco recomendadores distintos: (1) *Title recommender* - baseado em conteúdo, extrai *tags* do título do recurso formando um conjunto chamado *content based tags*; (2 e 3) *Title-to-tag* e *Tag-to-tag recommender* - dois recomendadores baseados em grafo que usam um algoritmos baseados em grafos direcionados de co-ocorrência de termos com objetivo de recomendar *tags*. Nestes grafos os vértices são representados pelas *tags* ou pelas palavras do título e as arestas quantificam o quão relacionado dois vértices estão. O grafo *title-to-tag* é usado no primeiro recomendador e o grafo *tag-to-tag* usado no segundo recomendador. Juntos geram a lista *resource related tags*. (4) *Resource profile recommender* - baseado em conteúdo, une a lista de *tags* já associadas ao recurso com a lista de *tags* relacionadas ao conteúdo deste recurso, juntas são chamadas de *resource related tags*; (5) *User profile recommender* - baseado em conteúdo, extrai *tags* do *profile* do usuário, chamada de *user related tags*. Para Lipczak e Milios [14], um recomendador de *tags* deve possuir três características:

1. Generalidade: os sistemas de recomendação de *tags* devem ser capazes de se adaptar automaticamente as características próprias dos ambientes de anotação colaborativa;

2. Adaptabilidade: as informações das novas anotações feitas pelos usuários devem ser usadas para melhorar a qualidade das *tags* recomendadas;

3. Eficiência: um sistema de recomendação de *tags* deve ser capaz de lidar com o grande volume de informação presente nos repositórios criados, de forma colaborativa, pelos usuários e devem ser eficientes para produzir resultados em tempo real.

Menezes *et al* [15], exploram a co-ocorrência de *tags* para criarem um recomendador automático de *tags* para páginas *Web*. Baseado em um conjunto inicial de *tags* (I_o) de um objeto o, o algoritmo LATRE, por eles proposto, extrai de um conjunto de treino somente regras de associação aplicáveis a o. A partir destas regras o algoritmo sugere as *tags* mais propensas a serem corretamente associado ao objeto o. Por fim, o conjunto de *tags* C_o que expande o conjunto I_o é derivado, sendo $I_o \cap C_o = \emptyset$.

Bischoff *et al* [5], questionam se *tags* podem ser usadas na busca de recursos. Para isso, eles avaliaram a sobreposição de um conjunto de *tags* obtidas do *Del.ici.ous*, do *Flickr* e do *Last.fm* [8] com o *log* de consultas da AOL. Eles obtiveram os seguintes resultados: $71,22\%$ das consultas eram compostas por pelo menos uma *tag* do

[8] http://www.last.fm

Del.ici.ous, enquanto $30,61\%$ das consultas eram completamente compostas por *tags* do *Del.ici.ous*. Para o *Flickr* e o *Last.fm* os números foram: $64,54\%$ e $12,66\%$ e 58.43% e 6% respectivamente. Com isso concluíram que a maioria das *tags* podem ser usadas na busca e que, na maioria dos casos, as *tags* usadas pelos usuários ao anotarem e ao buscarem um objeto são as mesmas. Porém, o simples fato de uma *tag* estar presente em uma consulta não necessariamente indica que o uso de *tags* como fonte de evidencia traria ganhos de qualidade em sistemas de RI. Um dos objetivos do nosso trabalho é estudar esse impacto mais a fundo por meio do estudo de técnicas de geração automática de *tags* e com experimentos que as incorporem como fonte de evidencia de relevância em sistemas de RI, tais como sistemas de busca e de classificação.

Uma das questores importantes a serem endereçadas é saber se as *tags* associadas a uma página são de boa qualidade. Um trabalho recente proposto por [19] visou criar métricas para avaliar a qualidade de um conjunto de *tags* associado a uma pagina. No total são exploradas oito métricas que avaliam desde a extensão (cardinalidade) de uma nuvem de *tags* até a popularidade das *tags* que a compõem. Em seu trabalho, Venetis *et al* afirmam que *tags* podem ser oriundas de palavras associadas a objetos (fotos, páginas *Web*, vídeo, etc) rotulados, de palavras extraídas do conteúdo textual desses objetos e dos rótulos pré-definidos das categorias usados para classificar documentos.

Admitindo S como uma nuvem de *tags*, as seguintes métricas foram utilizadas em [19] para avaliar a qualidade de S:

1. *Extent of* S: corresponde a cardinalidade de S, quanto maior a cardinalidade de S maior a sua cobertura, ou seja, um numero maior de tópicos será acessível por meio de S;

2. *Coverage of* S - corresponde a cobertura de S, quanto maior o número de documentos acessíveis via S maior a sua cobertura;

3. *Overlap of* S: diferentes tags em S podem estar associadas a um mesmo objeto anotado, a sobreposição captura a extensão desta redundância;

4. *Cohesiveness of* S: a coesão quantifica a similaridade dos documentos acessíveis por S, quanto maior o valor de coesão mais similares são os objetos acessíveis por S;

5. *Relevance of* S: a relevância mede o quão relevante S é para representar um determinado conjunto de documentos;

6. *Popularity of* S: a popularidade de uma tag $t \in S$ é especificada pela quantidade de objetos associados a t;

7. *Independence of S*: a independência quantifica o quão similar são os objetos rotulados por uma tag *t1* se comparados aos objetos rotulados por uma tag *t2*, ambas pertencentes a S;

8. *Balance of S*: considera a quantidade de documentos associados a cada tag $\in S$.

1.2 *Baseline*

Belem *et al*, em [4], propuseram várias novas heurísticas que expandem as estratégias usadas por [15] e [18] com o objetivo de recomendar *tags* para objetos *Web*. As novas heurísticas acrescentaram diferentes métricas aos modelos usados em [15] e [18]. A intensão de Belem *et al* é capturar quão precisamente um termo descreve o conteúdo de um objeto *Web* o a ser anotado. No total, foram propostas 8 novas estratégias de recomendação de *tags*.

Em seu trabalho, Belem *et al* exploram três tipos diferentes de fontes de informação: (1) a co-ocorrência de termos, (2) fontes textuais extraídas dos objetos, e (3) métricas que medem a relevância de *tags*. Como evidências textuais são usadas: as *tags* de o, o titulo de o e a descrição de o. As métricas, que medem a relevância das *tags*, usadas são: *Sum*, *Stability*, *TF*, *Entropy*, *IFF* e *AFS*, todas descritas em [4]. Pela combinação destas diferentes evidências é gera uma função de *rankig* que estima a relevância de uma *tag* t para o objeto o. Esta combinação é feita utilizando duas técnicas distintas de *learn-to-rank* (L2R): a GP (*genetic program*) e a RankSVM.

As bases de teste usadas no trabalho de Belem *et al* foram obtidas a partir dos sites *Last.Fm*, *YouTube* [9] e *YahooVideo*[10] . Em linhas gerais Belem *et al* definiram o problema de recomendação de *tags* da seguinte forma: Dado um conjunto de *tags* I_o previamente associado ao objeto o, e o conjunto de evidencias textuais, exceto as *tags*, $F_o = F_o^1, F_o^2, ..., F_o^n$, onde cada elemento de F_o^i é o conjunto de termos na evidência i do objeto o, é gerado um conjunto C_o de *tags* candidatas, de onde as k *tags* mais relevantes para o objeto o são obtidas.

Belem *et al* afirmam que suas novas heurísticas para recomendação de *tags* superam os resultados obtidos pelo melhor de seus *baselines*, produzindo ganhos em termos de precisão de até 40%, 32% de revocação e 62% de *MAP*.

Usaremos o trabalho de Belem *et al* como *baseline* na avaliação dos métodos de recomendação de *tags* aqui propostos (*agTag* e *agDirTag*). Verificaremos o quão bons são nossos resultados se comparados a este trabalho. Um diferencial a ser destacado é

[9] http://www.youtube.com/
[10] http://video.yahoo.com/

que em nossos métodos não é necessária a existência de um conjunto inicial de *tags* I previamente associado ao objeto *Web* a ser anotado, evitando problemas de *cold start*, ou seja, os métodos $agTag$ e $agDirTag$ são capazes de sugerir *tags* para páginas que nunca receberam anotações. Característica útil quando novos objetos são adicionados a *sites* sociais, cujo objetivo é compartilhar informações acerca de um objeto.

Esta proposta é constituída de quatro capítulos. No Capítulo 2 são apresentados os principais conceitos necessários para o entendimento do nosso trabalho. No Capítulo 3 estão descritos dois métodos, proposto por nós, para recomendação automática de *tags* para páginas *Web*. Neste mesmo capítulo são apresentados os experimentos realizados e discutidos seus respectivos resultados. Finalmente, no Capítulo 4 são apresentadas as conclusões e as direções futuras de trabalho.

Capítulo 2

Conceitos Básicos

2.1 Referencial Teórico

2.1.1 *Tags*

Tags são pequenas palavras que descrevem o conteúdo principal de um recurso ou de um objeto *Web* (páginas, *sites*, fotos, vídeos, músicas, etc). Elas podem ser obtidas do texto original do documento que se deseja anotar ou podem estar associadas ao tópico principal destes documentos [19]. Dado um recurso r, *tagging* é o processo onde usuários associam *tags* a r [9].

O *site Del.ici.ous*, em seu contexto, define *tags* como descritores (de tamanho um) associados a um *bookmark* que podem ser usados na organização e na busca destes últimos. Ainda afirma que as *tags* não formam uma hierarquia e que os próprios usuários são responsáveis por escolhe-las e associá-las aos objetos a serem anotados.

Já o *site Flickr* define *tag* como uma palavra-chave ou um rótulo de categoria. Neste contexto, as *tags* são usadas para encontrar fotos e vídeos que têm algo em comum. Esse ambiente permite que o usuário atribua até 75 *tags* a cada foto ou vídeo.

Existem duas formas de se associar uma *tag* a um documento: manual, onde o usuário conhecedor do conteúdo do documento define as *tags* que o descrevem; ou automático, onde sistemas (*tags recommenders*) exploram diversas fontes de informações para sugerir *tags* aos documento [12].

Gupta *et al* [9], identifica dez tipos de *tags* diferentes:

1. *Content-based tags*: usadas para identificar o real conteúdo dos recursos. Ex.: *Honda*, *batman*, *Lucene*.

2. *Context-based tags*: usadas para identificar o contexto no qual o objeto foi criado ou salvo. Ex.: *San Francisco*, *Golden Gate Bridge*, 2005-10-19.

3. *Attributed tags*: *tags* que são atributos inerentes do objeto e que não podem ser derivadas de seus conteúdos diretamente. Estas *tags* identificam o que ou sobre o que o recurso descreve, qualidade e características do recurso. Ex.: nome de autores como *Jeremy's blog* e *Clay Shirly*, *funny*, *stupid*.

4. *Ownership tags*: usadas para identificar o dono do recurso.

5. *Subjective tags*: usadas para expressar opiniões e sentimentos dos usuário. Ex.: *funny*, *cool*. Podem ser úteis na recomendação dos recursos.

6. *Organizational tags*: usadas para identificar coisas pessoais. Ex.: *mywork*, *mypaper*, *myhouse*. Também podem ser usadas como lembretes. Ex.: *to-read*, *to-review*.

7. *Purpose tags*: usadas para descrever o propósito de uma página com base em tarefas que podem ser realizadas por meio das mesmas. Ex.: *learn about latex*, *get recommendations*, *translate text*.

8. *Factual tags*: usadas para identificar as pessoas, lugares e conceitos de um recurso. São *tags* que a maioria das pessoas concorda em associar a um determinado recurso.Elas ajudam a descrever o objeto e a encontrar objetos relacionados. *Content-based, context-based* e *attribute tags* são consideradas *Factual tags*.

9. *Personal tags*: usadas para organizar as recursos dos usuários. São *tags* que expressam auto-referência, propriedade, organização de tarefas.

10. *Self-referential tags*: usadas para referenciar os próprios recursos. Ex.: uma imagem do *Flickr* que explica como usá-lo. As *tags* associadas a esta imagem são exemplos de *self-referential tags*.

11. *Tag Bundles*: são *tags* usadas para definir *tags*. Neste caso, usuários podem usar URLs para anotar outras URLs. Ex.: uma programa em C pode ser anotado com http://www.microsoft.com.

2.1.2 Nuvem de *Tags*

Nuvens de *tag* são representações visuais de documentos [19]. Elas são compostas por palavras, tipicamente *tags*, cujas formatações textuais (cor, espessura e tamanho

da fonte) são usadas para expressar o grau de importância (popularidade) destas *tags* dentro da nuvem [17]. As nuvens se tornaram comuns devido à popularização dos *sites* sociais que permitem a anotação, o armazenamento e o compartilhamento de diferentes recursos por diversos usuários. Alguns exemplos de *sites* sociais são: *Flickr* usado para compartilhar fotos, *Del.ici.ous* usado para compartilhar *bookmarks* e *LibraryThining*[11] que permite o compartilhamento de opiniões pessoais acerca de diversos livros. Nestes *sites* as nuvens de *tag* sumarizam a coleção de recursos exibindo as *tags* mais comumente utilizadas para descrever os recursos, ou seja, as *tags* mais populares [19]. Porém, o uso das nuvens de *tag* não se limita apenas aos *sites* sociais, estas já estão frequentes em páginas pessoais, em páginas comercias, em *blogs* e em *sites* de máquinas de busca. Nestes cenários as *tags* são extraídas do conteúdo textual das páginas e são ordenadas de acordo com suas respectivas frequências nos documentos [19]. Rivadeneira *et al* [17], explicam que além de sumarizar o conteúdo de uma coleção, as nuvens de *tag* podem ser usadas em:

- Busca: através das *tags* que formam uma nuvem os usuários podem localizar documentos acerca de um tópico específico ou ainda, localizar documentos relacionados ao tópico desejado.

- Navegação: as *tags* presentes nas nuvens são frequentemente *hyperlinks* que podem ser usados, como índices, por usuários para navegar pelos recursos de seu interesse.

- Formação de opiniões: observando o conjunto de *tags* de uma nuvem, o usuário poderá formar opiniões acerca da entidade (pessoa, empresa, instituição) associada a nuvem.

- Reconhecimento Correspondência: usuários podem usar nuvens de *tags* pessoais para determinar, por exemplo, qual dos dois José Silva é o pesquisador de informática que participou do projeto XYZ, pois o conjunto de *tags* associado a cada José Silva representa os interesses e as experiências particulares dos indivíduos.

Kuo *et al* [12] destacam que recentemente as nuvens de *tag* são opções importantes de *interface*, pois conseguem sumarizar de forma visual um grande conteúdo de informação.

[11] http://www.librarything.com/

2.1.3 Métricas de Avaliação para Sistemas de RI

Existem várias métricas usadas para estimar a qualidade dos resultados dos sistemas de RI em geral. Dentre elas, as mais utilizadas são Precisão (*precision*) e a Revocação (*recall*) [3]. Considerando N um conjunto de documentos relevantes e A um conjunto de documentos respondidos pelo sistema de RI a ser avaliado, pode-se determinar os seguintes conceitos:

- **Precisão (P)**: é a fração de documentos recuperados que podem ser considerados relevantes [3]. A Precisão pode ser definida pela Equação 2.1:

$$P = \frac{|N \cap A|}{|A|} \tag{2.1}$$

- **Revocação (R)**: é a fração de documentos relevantes que foi recuperada pelo sistema [3]. A Equação 2.2, define a revocação:

$$R = \frac{|N \cap A|}{|N|} \tag{2.2}$$

Onde, $|A|$ corresponde ao número de documentos no conjunto A, $|N|$ corresponde ao número de documentos no conjunto N e $|N \cap A|$ corresponde ao número de documentos pertencentes a interseção dos conjuntos N e A.

2.1.4 *N-grams*

De acordo com [8], *n-grams* é uma subsequência de n itens obtidos de uma sequência maior. Tais itens podem ser representados por letras, sílabas ou palavras de um texto, por exemplo. Neste trabalho os itens utilizados pra formar as subsequências são palavras e as subsequências formadas serão chamadas de termos. O valor de n representa a quantidade máxima de palavras utilizadas para compor um termo. Utilizaremos n igual a quatro, assim, termos compostos por uma, duas, três e quatro palavras serão utilizados no processo de identificação de tópico. Como exemplo, o texto *"Universidade Federal do Amazonas"* originaria os seguintes *n-grams*:

- *Uni-grams*: *"Universidade"*, *"Federal"*, *"do"*, *"Amazonas"*
- *Bi-grams*: *"Universidade Federal"*, *"Federal do"*, *"do Amazonas"*

- *Tri-grams*: *"Universidade Federal do"*, *"Federal do Amazonas"*
- *Quad-grams*: *"Universidade Federal do Amazonas"*

Os *n-grams* serão utilizados com o propósito de obter do texto de uma página *Web* os possíveis termos descritores desta. No exemplo acima, 10 termos se tornariam candidatos a descritores.

A técnica *n-grams* é frequentemente utilizada nos sistemas da área de RI. Ela será usada com objetivo de gerar novos termos para compor o conjunto de palavras candidatas a *tag* de uma página, a aplicação desta técnica é justifica, pois alguns termos isolados não possuem sentido completo.

2.1.5 Algoritmos Genéticos

Algoritmos Genéticos são métodos de busca e otimização inspirados nos conceitos da teoria de seleção natural das espécies. Os sistemas desenvolvidos a partir deste princípio são utilizados para procurar soluções de problemas complexos ou que possuam espaço de busca muito grande. Estes algoritmos são baseados nos processos genéticos (hereditariedade, mutação, seleção natural e cruzamento) de organismos biológicos para procurar soluções ótimas ou aproximadas do problema [13]. Para tanto, deve-se adequar o problema a ser resolvido aos requisitos exigidos por um AG, dessa forma três fases distintas de adequação do problema devem ser destacadas:

1. Codificação de cada possível solução do problema em uma estrutura chamada cromossomo;
2. Definição das configurações genéticas utilizadas: taxa de mutação, taxa de cruzamento, método de seleção, tamanho da população e número de gerações usadas no processo de evolução do AG.
3. Definição da função objetivo, que tem por finalidade avaliar o grau de adequação de cada cromossomo como solução do problema.

A utilização de algoritmo genético no método de sugestão de *tags*, proposto por nós, visa à descoberta dos pesos a serem associados a cada uma das evidências utilizadas. Os pesos (valores entre 0 e 1) serão utilizados em uma equação linear que define o grau de importância de um termo t para um documento D. Os AGs são boas técnicas utilizadas para atacar problemas de busca com espaço de busca intratavelmente grandes e que não podem ser resolvidos por técnicas tradicionais, como por exemplo a força bruta onde todas as possíveis soluções devem ser avaliadas [13].

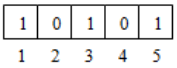
2.1.5.1 Codificação dos Cromossomos

A codificação dos cromossomos é fundamental na modelagem o algoritmo genético, ela consiste em uma maneira de traduzir a informação do problema a ser resolvido em uma forma viável a ser tratada pelo AG [13].

Os cromossomos representam possíveis soluções do problema e podem ser vistos como um ponto, do espaço de busca, candidato a solução. Eles devem ser codificados de acordo com as características do problema a ser resolvido. Cada cromossomo é composto por vários genes e cada gene representa um aspecto distinto da solução.

Os cromossomos possuem diferentes formas de serem representados, entre elas: binária, inteira ou real. A essa representação se dá o nome de alfabeto do AG [13]. De acordo com a classe de problema que se deseja resolver pode-se usar qualquer um dos tipos.

Figura 2.1. Exemplo de codificação de um cromossomo com 5 genes e alfabeto binário.



2.1.5.2 Operadores Genéticos

Os operadores genéticos mais conhecidos e utilizados nos algoritmos genéticos são os de seleção, cruzamento (*crossover*) e de mutação [13].

- **Seleção**: este operador seleciona cromossomos da população para a realização da reprodução. Quanto maior a aptidão do cromossomo maior é a chance dele ser escolhido para reprodução.

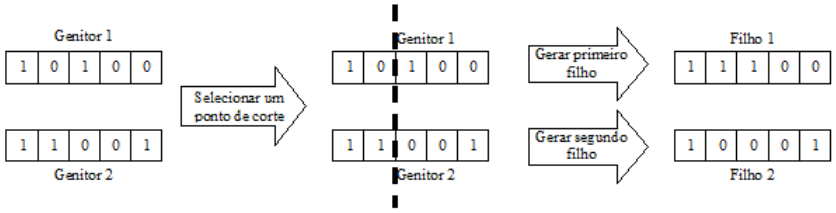
  O método de seleção de pais deve ser semelhante ao mecanismo de seleção natural que atua sobre as espécies biológicas, em que pais mais aptos geram mais filhos e pais menos aptos geram menos filhos [13]. Consequentemente deve-se privilegiar os indivíduos mais aptos, sem desprezar completamente os de aptidão inferior, pois, estes podem ter características genéticas que sejam favoráveis à criação de um indivíduo que representa a melhor solução para o problema. Por outro lado, se apenas os melhores indivíduos se reproduzirem ocorrerá um efeito chamado de convergência genética.

  A convergência genética ocorre quando a população se compõe por indivíduos cada vez mais semelhantes, acarretando a falta de diversidade o que impede a evolução satisfatória da população [13].

O método da Roleta é uma maneira de selecionar indivíduos, onde cada indivíduo possui uma fatia da roleta proporcional à sua adaptação. A cada giro da roleta um indivíduo é selecionado, tendo maior chance aqueles que possuem as maiores fatias, sem deixar de lado a diversidade dos menos adaptados [13]. Outras formas de seleção podem ser aplicadas dependendo do problema a ser tratado.

- **Cruzamento (*Crossover*)**: operador genético que cria novos indivíduos através da combinação das características de outros dois indivíduos. Este processo é ilustrado na Figura 2.2, onde a solução está codificada no alfabeto binário.

Figura 2.2. Descrição da operação genética *crossover*.



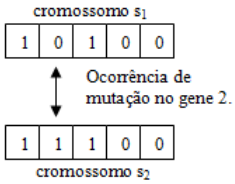
O funcionamento do operador genético de cruzamento consiste em: (1) selecionar os cromossomos genitores; (2) escolher, aleatoriamente, o ponto onde ocorrerá o corte para a realização do cruzamento; (3) separar as características genéticas dos cromossomos genitores em duas partes (uma a esquerda e outra a direita do ponto de corte); (4) gerar o primeiro filho, que será composto pela parte esquerda do primeiro "pai" e pela parte direita do segundo "pai"; e (5) gerar o segundo filho, que será composto pela parte direita do primeiro "pai" e pela parte esquerda do segundo "pai" [13]. A realização do cruzamento garante a troca de informações genéticas entre diferentes e possíveis soluções.

- **Mutação**: é um operador unário que cria novos indivíduos através da modificação aleatória dos valores contidos em um ou mais genes de um cromossomo.

  Ao operador de mutação é associada uma probabilidade baixa de ocorrência, caso contrário o funcionamento do AG se parecerá com uma técnica chamada *random walk*, na qual a solução é determinada de forma aleatória [13]. Quando a probabilidade atinge um gene em questão, então seu valor é aleatoriamente alterado por outro pertencente ao domínio válido [13]. A mutação garante a diversidade das características dos indivíduos da população e permite que sejam introduzidas características que não estavam presentes em nenhum dos indivíduos [13]. O valor da probabilidade de ocorrência de mutação (taxa de mutação) é definido como

parâmetro do AG. A Figura 2.3 demonstra graficamente a ocorrência de mutação no gene 2 do cromossomo, cuja solução está codificada no alfabeto binário.

Figura 2.3. Descrição da operação genética mutação.



2.1.5.3 Função Objetivo

A função objetivo tem por finalidade determinar a qualidade de um indivíduo (cromossomo) como solução do problema, isto é, ela retorna um valor numérico que reflete quão bons os parâmetros representados no cromossomo resolvem o problema em questão [13]. A função objetivo deve refletir os objetivos a serem alcançados na resolução do problema. Na Seção 3.1.1.2 explicaremos a função objetivo adotada no problema a ser resolvido pelo AG.

2.1.6 Validação Cruzada

A validação cruzada é uma técnica frequentemente utilizada em aprendizado de máquinas (*machine learning*). Ela consiste em dividir um conjunto de dados D em n subconjunto D_i. Com essa divisão um dado algoritmo pode ser executar n vezes, cada vez usando um conjunto de treino diferente (D - D_i) e o teste do algoritmo pode ser feito com o subconjunto D_i [6].

O conjunto de treino ainda é subdividido em um subconjunto de dados para validação (D_v) e um subconjunto de dados para estimação (D - D_i - D_v). A ideia é utilizar o conjunto de treino para avaliar o desempenho dos indivíduos candidatos à solução do problema e assim, escolher o melhor. O subconjunto de treino permite selecionar o indivíduo e o subconjunto de validação permite validar o indivíduo escolhido como solução. Já com o conjunto de testes é verificada a generalização do modelo.

Neste trabalho, aplicaremos a técnica de validação cruzada no treino, validação e teste do algoritmo genético utilizado para determinar a combinação das diferentes fontes de evidências usadas nos métodos de recomendação automática de *tags* para uma páginas *Web*.

Capítulo 3

Recomendação Automática de *Tags*

Como citado no Capítulo 1, um de nossos objetivos é propor e avaliar métodos de recomendação automática de *tags* para páginas *Web*. A seguir serão expostos dois métodos já desenvolvidos e avaliados por nós. Ambos exploram informações provindas de diversas fontes de evidências textuais (extraídas de uma página p) e informações extraídas da hierarquia de categorias de um diretório para gerarem conjuntos de *tags* a serem recomendados a p. São eles:

1. ***agTag*** - método que recomenda um conjunto de novas *tags* para uma página p através da combinação de várias fontes de evidências textuais de p e métricas que medem relevância dos termos de p.

2. ***agDirTag*** - método que expande o resultado do *agTag* agregando informações oriundas da hierarquia de categorias do diretório *ODP* [12] (*Open Directory Project*) para sugerir um novo conjunto de *tags* para p.

Para que os métodos fossem desenvolvidos e avaliados foi necessária a criação de uma base de referência. O objetivo desta base é servir como um repositório de documentos que possam ser utilizados nos experimentos do processo de recomendação automática de *tags*. Para compô-la foram obtidas páginas pertencentes ao *site Del.ici.ous*. Apenas uma restrição foi imposta: o idioma das páginas deve ser o inglês, pois pesquisas feitas no diretório *ODP* comprovam que existe um maior número de páginas classificadas neste idioma se compararmos ao número de páginas classificadas em outros idiomas [1]. A escassez de informações no diretório utilizado é observada como um ponto negativo, podendo influenciar negativamente na avaliação do método *agDirTag*.

[12] http://www.dmoz.org/

Cada página p pertencente a base de referência é representada por uma tupla $G = <TextoHtml, TextoÂncora, TagsDescritoras>$, onde *TextoHtml* corresponde ao conteúdo textual da página p incluindo as *tags* de marcação HTML, *TextoÂncora* corresponde a concatenação dos textos de âncora dos *links* que referenciam p e *TagsDescritoras* corresponde ao conjunto de *tags* pré-associadas a p obtidas do *site Del.ici.ous*. É importante observar que as *tags* descritoras de uma página p foram geradas por diferentes usuário que expressaram suas opiniões a respeito do conteúdo de p. Assim, cada página pertencente à base de referência possui um conjunto de *tags* pré-definidas e que será utilizado posteriormente na geração de um modelo que identificará novas *tags* para páginas *Web*.

3.1 Método *agTag*

O método *agTag* combina diversas fontes de evidências para cumprir seu objetivo: sugerir o conjunto *tags_ agTag* de boas *tags* para p. Neste método dois tipos de fontes de evidências foram exploradas: (1) evidências textuais, como: o texto completo das página, os textos destacados por algumas *tags* HTML e a concatenação dos textos de âncora de *links* que referenciam p; e (2) métricas que medem a relevância de um termo para uma página, neste caso o valor de TF×IDF.

Tendo em mente que o usuário desenvolvedor da página tende a destacar suas principais frases e palavras, e ainda que estas possam conter boas *tags* para p, o conteúdo das seguintes *tags* HTML foram selecionadas como fonte de evidência: <H1>, <H2>, <Bold>, <Strong> e <Title>. Outra fonte de evidência utilizada pelo método é o valor de TF×IDF de cada termo t candidato a *tag* da página p. O objetivo da utilização desta fonte de evidência é mensurar a importância de t como *tag* de p. Além das *tags* HTML e do valor de TF×IDF, também foi utilizado como fonte de evidência a concatenação dos textos de âncora de *links* que referenciam a página p,explorando desta forma as múltiplas visões existentes acerca do conteúdo de p.

Em linhas gerais, o método *agTag* está dividido em duas etapas distintas:

1. *Identificação do Modelo*: etapa supervisionada que consiste em identificar um modelo genérico a ser utilizado na recomendação de *tags* para páginas *Web*. Uma vez identificado o modelo, este pode ser usado na recomendação de *tags* para novas páginas *Web*.

2. *Recomendação de Tags*: etapa não supervisionada que consiste em sugerir um conjunto de *tags* para uma página *Web* qualquer p.

3.1.1 Identificação do Modelo

O objetivo desta etapa é determinar os pesos a serem associados a cada uma das seis evidências utilizadas pelo método *agTag*. Os pesos serão utilizados na equação linear 3.1, utilizada para calcular a importância de um termo t para um documento D. O processo de descoberta dos pesos deve ser automático, evitando as suposições humanas, para isso exploramos a técnica de algoritmos genéticos.

$$\begin{aligned} Imp(t,D) &= [\alpha \times eH1(t,D)] + [\beta \times eH2(t,D)] + [\gamma \times eBold(t,D)] \\ &\quad + [\delta \times eTitle(t,D)] + [\epsilon \times vTF \times IDF(t)] \\ &\quad + [\theta \times eAncTex(t,D)] \end{aligned} \tag{3.1}$$

Na Equação 3.1 o valor de $Imp(t,D)$ representa a importância de um termo t como *tag* para o documento D, D é uma página *Web*. $eH1(t,D)$ corresponde a frequência com que o termo t ocorre entre as *tags* <H1> e </H1> no documento D. $eH2(t,D)$ corresponde a frequência com que o termo t ocorre destacado pela *tag* <H2> no documento D. $eBold(t,D)$ corresponde a frequência com que o termo t ocorre destacado pelas *tags* HTML <Bold> e <Strong> no documento D. $eTitle(t,D)$ corresponde a frequência com que o termo t ocorre entre as *tags* <Title> e </Title> em D. *vTF×IDF(t)* corresponde ao valor de TF×IDF relativo do termo *t*. E, $eAncText(t,D)$ corresponde a frequência com que o termo t ocorre na concatenação dos textos de âncora dos *links* que apontam para D. Já α, β, γ, δ, ϵ e θ representam os pesos associados as evidências H1, H2, Bold, Title, TF×IDF e texto de âncora respectivamente, sendo seus valores obtidos através do uso de algoritmos genéticos.

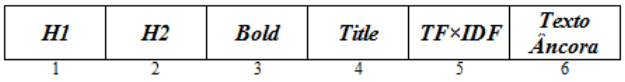
O uso de algoritmos genéticos nesta etapa do processo de recomendação de *tags* a páginas *Web* visa aperfeiçoar a escolha dos pesos associados a cada evidência utilizada, pois o espaço de solução do subproblema em questão é muito grande (conjunto dos números reais). Logo, temos um problema de busca no qual um AG específico será utilizado para solucioná-lo. Objetivando o uso de um AG na descoberta dos pesos das evidências, deve-se adequar o subproblema, de descoberta de pesos, aos requisitos de um AG, dessa forma quatro fases distintas de adequação do problema são destacadas:

3.1.1.1 Planejamento dos Cromossomos

Como visto na no Seção 2.1.5.1, os cromossomos representam possíveis soluções do problema e devem ser modelados de acordo com as características do problema a ser resolvido. Dentre os possíveis tipos de representação utilizamos a real, ou seja, cada

gene pertencente ao cromossomo representará um número real compreendido entre 0 e 1. Estes valores corresponderão aos valores de α, β, γ, δ, ϵ e θ, que representam, respectivamente, os pesos atribuídos as evidências H1, H2, Bold, Title, TF×IDF e texto de âncora, aplicados a Equação 3.1 que tem o objetivo de identificar as melhores *tags* de uma página *Web*. A Figura 3.1 demonstra graficamente o cromossomo utilizado na resolução do subproblema. Cada cromossomo é composto por seis genes que representam os pesos atribuídos a cada uma das seis evidências utilizadas no processo de recomendação de *tags*. Assim, o primeiro, segundo, terceiro, quarto, quinto e o sexto gene representarão os pesos atribuídos às evidências H1, H2, Bold, Title, TF×IDF e ao texto de âncora respectivamente.

Figura 3.1. Representação do cromossomo utilizado na solução do subproblema de identificação dos pesos das evidências utilizadas pelos métodos de recomendação de *tags*.



3.1.1.2 Definição da Função Objetivo

A função objetivo tem por finalidade avaliar o grau de adequação de cada indivíduo (cromossomo) como solução do problema, sendo associado a cada cromossomo um valor de aptidão gerado por esta função. A função objetivo, aqui utilizada, reflete o objetivo do subproblema em questão, ou seja, o de encontrar os melhores pesos para as seis evidências utilizadas na recomendação de um conjunto de *tags* para uma página *Web*, através da maximização de seu valor.

A função objetivo utilizada no AG foi a Medida F_1 (*F-measure* ou Média Harmônica). Ela é útil quando se deseja combinar os valores de Precisão P e Revocação R (Seção 2.1.3) em um único valor de medida de qualidade [3]. A medida F_1 pode ser definida como mostra a Equação 3.2.

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{3.2}$$

A função F_1 assume valores entre 0 e 1. Quando utilizada na avaliação de sistemas de busca, o resultado 0 indica que nenhum documento relevante foi recuperado e o resultado 1 indica que o sistema conseguiu recuperar todos os documentos relevantes com precisão máxima. Consequentemente, o máximo valor de F_1 pode ser interpretado como o melhor resultado da combinação de precisão e revocação [3].

O valor de F_1 foi utilizado como função objetivo do AG aqui proposto, pois nosso objetivo é obter um conjunto T de *tags* para a página *Web* p que contenha as seguintes características: (1) T deve conter boas *tags* para p, maximizando a precisão P e (2) as boas *tags* que definem p devem está presente em T, maximizando a revocação R. Portanto, nossa intenção é maximizar o valor de F_1.

Contudo, a fase de treino do AG também foi usada para a descoberta de um limiar c utilizado na classificação das possíveis *tags* de p como sendo boas ou não. Dezoito valores entre 0 e 1 foram testados a fim de encontra um bom limiar de classificação. Os valores 0 e 1 representam os extremos de retorno da função F_1 (função objetivo utilizada pelo AG). Na Seção 3.3.2 será explicado a forma de escolha dos valores de limiar.

O cálculo do valor de F_1 para uma determinada página p, pertencente a base de referência, dada uma possível solução s_i $(1 \leq i \leq 500)$ do problema, é feito da seguinte forma:

1. A partir da página p e da tupla G= <*TextoHtml, TextoÂncora, TagsDescritoras*> que a representa é gerado um conjunto T de possíveis *tags* de p, conforme a Figura 3.2 e detalhamento nas Seções (3.1.2.2), (3.1.2.3), (3.1.2.4) e (3.1.2.5).

2. A partir do elemento *TagsDescritoras* da tupla G, obtemos o conjunto de *tags* pré-definidas de p. Deste conjunto, são eliminadas as *tags* de freqüência um, pois foi constatado, por amostragem, que muitas delas são *tags* ruidosas, ou seja, *tags* não relevantes para o contexto da página ou ainda *tags* de baixa qualidade. Desta forma, gera-se um novo conjunto I de *tags* pré-definidas de p. O conjunto I_p é utilizado inicialmente como um conjunto ideal de *tags* para p. Como exemplo de *tags* ruidosas podemos destacar: *"*"*, *"-"*, *"all"*, *"educator's_ blogs"*, *"analysis"*, *"form"*.

3. O próximo passo é calcular para cada possível *tag* t pertencente a T o valor de $Imp(t, p)$ (conforme Equação 3.1), ou seja, a importância de t para p. Os pesos associados a cada uma das evidências utilizadas na equação 3.1 são representados pelos genes do cromossomo s_1.

4. Em seguida, são selecionadas do conjunto de *tags* T, apenas aquelas que possuem o valor de $Imp(t, p)$ acima de um limiar c_j $(1 \leq j \leq 18)$. Considerando este último conjunto como sendo a resposta do sistema e o conjunto I_p como resposta ideal, foi calculado o valor de F_1 para a página p e limiar c_j.

5. Os passos 1 a 4 são executados para cada conjunto de páginas de treino do AG e para todos os dezoito limiares testados.

Por fim, para cada limiar c_j testado é calculada a média dos valores de F_1 das páginas utilizadas no treino da AG. O maior valor médio encontrado de F_1 identifica a solução s_i ótima encontrada e o melhor valor c_j de limiar a ser utilizado. A solução ótima s_i contém aos valores dos pesos a serem associados a cada uma das evidências da Equação 3.1.

3.1.1.3 Definição dos Parâmetros Genéticos

Além de planejar os cromossomos e definir a função objetivo, algumas configurações genéticas devem ser definidas objetivando a melhora do desempenho do AG. Na solução do subproblema de definição dos pesos das evidências foram utilizados diversos parâmetros de configuração no algoritmo genético. Abaixo estão expostas as configurações genéticas utilizadas no AG e sugeridos em [13]:

- Taxa de crossover: 70% dos cromossomos.
- Taxa de mutação: 3% dos genes.
- Tamanho da população: 500 indivíduos.
- Seleção de indivíduos: Método da Roleta (Seção 2.1.5.2).
- Criação da população: a população inicial foi criada com valores aleatórios
- Gerações: 10 e 20 gerações

3.1.1.4 Processo de Evolução

O algoritmo 1 descreve o processo evolutivo do AG utilizado. Para cada possível solução do problema (cromossomo) s_i ($1 \leq i \leq 500$) são gerados dezoito conjunto de *tags*, $T_{Imp(t,p)>c_j}$ ($1 \leq j \leq 18$), onde apenas aquelas *tags* que possuíssem o valor de $Imp(t,p)$ acima de um limiar c_j pertencem a $T_{Imp(t,p)>c_j}$.

Para cada limiar c_j testado é calculada a média dos valores de F_1 das páginas utilizadas no treino da AG. O maior valor médio encontrado de F_1 identifica a solução s_i ótima e o melhor valor c_j de limiar a ser utilizado. A solução ótima s_i contém aos valores dos pesos a serem associados a cada uma das evidências da Equação 3.1.

Algorithm 1 Processo evolutivo da AG

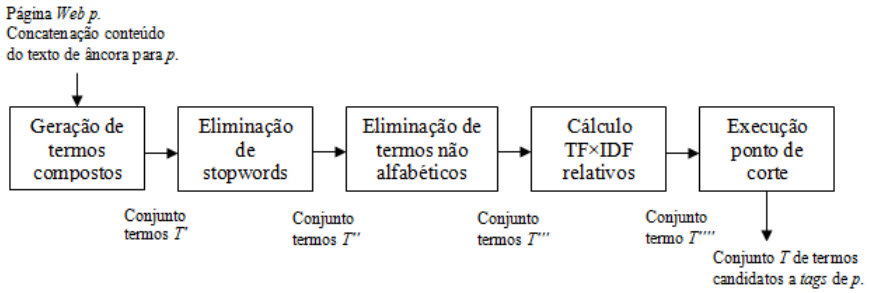
```
Let Limiar[] = {0.01, 0.015, 0.02, 0.025, ..., 0.095}
for all s_i ∈ S do
    α ← s_i[1]
    β ← s_i[2]
    γ ← s_i[3]
    δ ← s_i[4]
    ϵ ← s_i[5]
    θ ← s_i[6]
    for cada página p ∈ ao conjunto de treino do AG do
        gera T_p conforme Figura 3.2
        gera I_p conforme Seção 3.1.1.2
        for all t_i ∈ T_p do
            Calcula Imp(t_i, p)
        end for
        for all Limiar[j], 1 ≤ j ≤ 18 do
            Separa conjunto de tags cujo valor de Imp(t, p) ≥ Limiar[j], gerando
            T_{Imp(t,p)>c_j}
            Calcula valor F_1 conforme Equação 3.2
        end for
    end for
end for
Identifica melhor solução do problema
```

A finalização da fase de Identificação do Modelo ocorre ao final do ciclo de execução de treinos do AG, o cromossomo com maior valor de aptidão corresponde a melhor solução do subproblema de identificação dos pesos das evidências e os valores, por ele representado, serão aplicados a fórmula da Equação 3.1. Além disso, o valor de limiar associado a melhor solução é utilizado para classificar possíveis *tags* de uma página qualquer p como boas ou não.

3.1.2 Recomendação de *Tags*

O processo de sugestão de *tags* para p é subdividido em outras cinco fases, como mostra a Figura 3.2. Na primeira subfase é gerado um conjunto inicial T' contendo todos os termos candidatos a possíveis *tags* de p, nas subfases seguintes o conjunto T' passa por mudanças (exclusão de termos) gerando conjuntos intermediários que ao final do processo darão origem ao conjunto T_p de termos candidatos a *tags* de p. A seguir, será explicada cada subfase.

Figura 3.2. Subfases da etapa de recomendação de *tags*.



3.1.2.1 Fase 1: Geração de Termos Compostos

Utilizando-se da técnica de *n-grams*, descrita na Seção 2.1.4, foram gerados os termos para compor o conjunto de possíveis *tags* de p, a aplicação da técnica é justifica, pois alguns termos isolados não possuem sentido completo. O valor de n utilizado no método aqui exposto é quatro, assim as palavras compostas por um, dois, três e quatro termos, podem fazer parte do conjunto de possíveis *tags* de p. Uma restrição imposta durante esta etapa é que os novos termos gerados não podem começar ou terminar com *stopwords*, porém estas podem aparecer no meio da composição do termo. Como exemplo destacamos o termo: *"department of energy"*, onde a *stopword "of"* aparece sua na composição.

3.1.2.2 Fase 2: Eliminação de *Stopwords*

A eliminação de termos muito comuns como artigos, preposições, pronomes e algumas palavras como *e-mail*, *everything* e *few* se faz necessário com a finalidade de minimizar o tamanho do conjunto de possíveis *tags* de uma página. Tais termos muitas vezes não possuem significados relevantes que possam influenciar no resultado final do método. Com a eliminação de *stopwords* há uma diminuição no conjunto de possíveis *tags* de p. A lista de termos classificados como *stopwords* pode ser extensiva a alguns verbos, advérbios e adjetivos além dos artigos, preposições e conjunções [3].

3.1.2.3 Fase 3: Eliminação de Termos não Alfabéticos

Com o mesmo propósito justificado pela eliminação das *stopwords*, termos não alfabéticos como números, datas, códigos de endereçamento postal, números de telefone e endereços eletrônicos são eliminados.

3.1.2.4 Fase 4: Cálculo do TF×IDF relativo

O valor de TF (*Term Frequency*) de um termo qualquer t presente em um documento indica a freqüência com que t ocorre no documento, assim a freqüência de que cada t candidato a *tag* de p foi computada.

Para cálculo dos valores de *IDF*, a coleção utilizada foi a *Web*. Requisições foram feitas a máquina de busca *Google* com o objetivo de encontrar o valor aproximado do total de documentos indexados pela máquina e o valor aproximado do total de documentos que contêm cada termo t candidato a *tag* de p. Por este motivo o valor de *IDF* utilizado não é o real e sim o relativo à base indexada pela máquina de busca utilizada. Todos os valores de *IDF* necessários durante o processo de identificação de *tags* foram pré-calculados e armazenados em memória.

3.1.2.5 Fase 5: Execução dos Pontos de Corte

Ainda com a finalidade de diminuir o tamanho do conjunto de possíveis *tags* de p, eliminamos os termos cujo valor de *IDF* relativo seja superior a 16, pois, realizando experimentos e analisando os cálculos de *IDF*, percebemos que termos com altos valores de IDF representam um conjunto de termos que pode ser descartado do conjunto de possíveis *tags* para a página p, pois se tratam de termos extremamente raros.

A finalização do processo de recomendação de *tags* para uma página *Web* p, dá-se quando o método *agTag* seleciona do conjunto de *tags* T as *tags* cujo valor de $Imp(t, D)$ seja superior a um limiar c_j. Os pesos associados a cada evidência da equação $Imp(t, D)$ e o valor do limiar c_j foram identificados na seção 3.1.1. Nomearemos o conjunto de *tags* sugeridas pelo método *agTag* para uma página p como *tags_ agTag*.

3.2 Método *agDirTag*

O segundo método proposto e avaliado é o *agDirTag*. Ele expande o resultado do método *agTag* explorando informações oriundas da hierarquia de categorias do diretório *ODP* para sugerir um novo conjunto de *tags* para uma página p.

Como dado de entrada a este método é fornecido o conjunto *tags_ agTag* de *tags* sugeridas pelo método *agTag* para uma página p. Como dado de saída é sugerido um novo conjunto *tags_ agDirTag* de *tags* para p. Para determinar o conjunto *tags_ agDirTag* de novas *tags* para p o método *agDirTag* utilizada informações da estrutura hierárquica de um diretório.

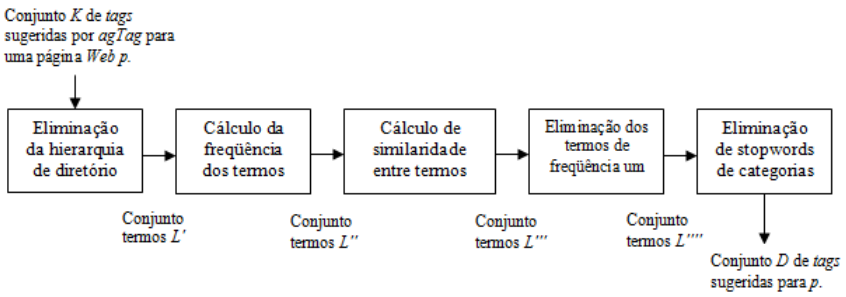
Inicialmente, todos os elementos $k_i \in tags\_agTag$ da página p são submetidos ao serviço de busca do diretório, em seguida as 50 primeiras categorias associadas às 50 primeiras respostas retornadas pela submissão são obtidas e geram um lista inicial L com todas as 50 categorias obtidas. O método então, utiliza os termos presentes na lista de categorias para associar novas *tags* a página *Web* p. Os termos da lista passam por um processo de filtragem que visa escolher quais os termos mais adequados para compor o conjunto $tags\_agDirTag$ de p por meio de heurísticas de eliminação.

O serviço de diretório utilizado nos exemplos e experimentos aqui citados foi o *ODP*. O diretório *ODP*, também denominado *DMOZ* (*Directory Mozilla*), consiste em uma coleção de *sites* organizados por tópicos em uma estrutura hierárquica de categorias. O *ODP* é o maior e mais abrangente diretório editado por seres humanos na *Web*, ele é mantido por um grupo de editores voluntários de todo o mundo que avalia manualmente as informações a serem incluídas no diretório [1]. As páginas da *Web* avaliadas por esses editores são classificadas em diversas categorias que compõem a hierarquia do diretório.

Como resposta a uma busca, o diretório *ODP* retorna uma lista ordenada de páginas referente ao assunto buscado e a categoria a qual estas últimas pertencem. Estas categorias foram exploradas como fonte de informação adicional pelo método *agDirTag* com a finalidade de expandir o conjunto de *tags* já recomendadas pelo método *agTag*.

O processo de filtragem pelo o qual da lista L é submetida visa encontrar as *tags* que melhor descrevem o conteúdo de p. Esta processo é feito em cinco etapas distintas conforme mostra a Figura 3.3.

Figura 3.3. Etapas do processo de filtragem da lista de categorias L, objetivando a sugestão de *tags* para uma página *Web*



3.2.1 Etapa 1: Eliminação da hierarquia de diretório

Na primeira etapa do processo de filtragem da lista L é feita a eliminação da estrutura hierárquica das categorias, isto é, a relação pai/filho utilizada pela taxonomia do diretório deixa de ser relevante e os termos que antes compunham a categoria passam a ser tratados de forma individualizada, formando um conjunto L' de palavras candidatas a *tags* de p. O descarte da hierarquia se justifica, pois, conforme descrito em [10], a relação pai/filho utilizada pela taxonomia dos diretórios disponíveis pode ser de dois tipos: (1) "é um"; ou (2) "é parte de", o que impede a afirmação de que todas as palavras que compõem a hierarquia de categorias podem ser classificadas como boas *tags*. Como exemplo, suponha que a seguinte hierarquia de categorias foi identificada como sendo a hierarquia que melhor descreve o termo *"John Paul II"* (Papa): */top/Society/Religion_and_Spirituality/Christianity/Denominations/Catholicism/Saints/J/Blessed_John_Paul_II*, pode-se afirmar que uma página que discorre sobre *"John Paul II"* tenha como exemplo de boa *tag* a palavra *"Christianity"*, porém o mesmo não ocorre se tomarmos como exemplo de *tags* as palavras *"Top"* ou *"J"*.

3.2.2 Etapa 2: Cálculo da frequência dos termos

Na segunda etapa é computada a frequência com que cada termo l_i ocorre em L'. As frequências são atribuídas a cada l_i. O cálculo da frequência dos termos se justifica, pois em uma próxima etapa este valor será usado como ponto de corte. Como saída desta etapa temos um novo conjunto $L"$ de termos candidatos a *tags* de p.

3.2.3 Etapa 3: Cálculo de similaridade entre termos

A terceira etapa consiste em identificar a similaridade entre os termos presentes no conjunto $L"$ e termos pertencentes ao conjunto *tags_agTag* (conjunto de *tags* sugeridas pelo método *agTag* para p). Os termos presentes ao conjunto $L"$, cuja grafia coincida com ou contenha termos do conjunto *tags_agTag*, são considerados boas *tags* para a página p. As boas *tags* encontradas são retiradas do conjunto $L"$ e inseridas no conjunto *tags_agDirTag* (conjunto definitivo de *tags* recomendadas para p pelo método *agDirTag*). Como saída temos um novo conjunto $L"'$ de termos candidatos a *tags* de p.

3.2.4 Etapa 4: Eliminação dos termos de frequência um

Nesta etapa, são eliminados os termos de *L"'* que possuem frequência igual a 1, o que acarreta na diminuição da quantidade de termos candidatos a *tag* de *p*. A eliminação destes termos é feita, pois a maioria dos termos que possuem peso 1, após a execução da terceira etapa de filtragem, são termos considerados irrelevantes. Como saída desta etapa temos um novo conjunto de candidatas a *tags* de *p*: *L""*.

3.2.5 Etapa 5: Eliminação de *stopwords* de categorias

Na quinta, e última, etapa é feita a eliminação dos termos presentes na lista de *stopwords* de categorias. Termos considerados *stopwords* de categorias são aqueles de sentido amplo utilizados apenas para compor a hierarquia das categorias do diretório, que isolados não possuem sentido completo e sob as quais estão muitas outras subcategorias específicas, como por exemplo: *"top"*, *"issue"*, *"reference"*, *"by_region"*. A lista de *stopwords* de categorias foi feita analisando a estrutura hierárquica do diretório *ODP*.

Como saída desta etapa é gerado o conjunto final de *tags* para *p*, acrescentando ao conjunto *tags_agDirTag* os elementos restantes no conjunto *L""*. Logo, como resposta final do método *agDirTag* temos o conjunto *tags_agDirTag* de *tags* para *p*.

3.3 Experimentos Realizados

Experimentos foram realizados para averiguar a eficiência dos métodos expostos. A seguir serão apresentados os experimentos realizados, discutidos os objetivos de cada experimento, as configurações utilizadas, as formas de execução e os resultados obtidos.

3.3.1 Composição da base de referência

Como citado no início deste capítulo, foi necessária a formação de uma base de referência que servisse como um repositório de documentos a serem utilizados nos experimentos dos métodos de recomendação de *tags* aqui propostos. Foram selecionadas randomicamente 112 páginas da base coletada por Menezes *et al* em [15]. Menezes *et al* coletaram, em outubro de 2009, 560mil objetos (*bookmarks*) da página *Recent* do *site Del.ici.ous*. Esta página apresenta em ordem cronológica os objetos do *site* que sofreram recente alteração. A partir desses objetos foram coletados o conteúdo de suas respectivas páginas (*bookmarked pages*) e o conjunto de *tags* associadas a cada objeto (*top tags*).

Para compor a base de referência todas as páginas devem ser representadas por uma tupla G, desta forma: $G = <TextoHtml, TextoÂncora, TagsDescritoras>$. Onde, o elemento *TextoHtml* representa o conteúdo textual com *tags* HTML de uma página p, o elemento *TextoÂncora* representa a concatenação dos textos de âncora dos *links* que apontam para p e o elemento *TagsDescritoras* representam o conjunto de *tags* associadas a cada página p através do *site Del.ici.ous*.

Objetivando a exploração exaustiva da base de referência, optou-se por utilizar a técnica de validação cruzada (Seção 2.1.6) nos experimentos realizados. Logo, a base de referência foi dividida em quatro subconjuntos (4 *folds*), cada um contendo 28 páginas. Do total de páginas que compunham a base 80% foram utilizadas para treino e 20% foram utilizadas para a teste.

3.3.2 Experimento 1: Identificação do limiar de classificação de *tags* e Obtenção dos pesos das evidências

Este experimento foi realizado com o objetivo de determinar os melhores pesos das seis fontes de evidências utilizadas pelos métodos de recomendação de *tags* e o melhor valor de limiar utilizados para classificar uma *tag* candidata como sendo boa ou não.

Os pesos, aqui definidos, foram aplicados na Equação 3.1 que quantifica a importância de um termo t como *tag* para uma página p. A Seção 3.1.1.3 enumera os vários parâmetros genéticos utilizados durante a execução do AG que definirá os pesos das evidências. O uso desta variedade de configurações tem como objetivo obter o melhor resultado para a solução do problema, ou seja, maximizar o valor da função objetivo utilizada. A função objetivo utilizada para avaliar o grau de aptidão de cada possível solução (cromossomo) foi a medida F_1 explicada na Seção 3.1.1.2.

Durante a fase de treino do AG também aprendemos o valor de um limiar utilizado para classificar uma *tag* como boa para a página p. Sabendo que o valor da função $Imp(t, D)$ varia entre 0 e 1, foram testados dezoito valores distintos para o limiar de classificação (c): 0,1; 0,15; 0,2; 0,25; 0,3; 0,35; 0,4; 0,45; 0,5; 0,55; 0,6; 0,65; 0,7; 0,75; 0,8; 0,85; 0,9 e 0,95. Porém, os experimentos mostraram que a maioria dos valores obtidos pela função $Imp(t, D)$, para a coleção, concentra-se entre 0 e $0,1$. Logo, mudamos os possíveis valores do limiar c testados para: 0,01; 0,015; 0,02; 0,025; 0,03; 0,035; 0,04; 0,045; 0,05; 0,055; 0,06; 0,065; 0,07; 0,075; 0,08; 0,085; 0,09 e 0,095

Definidos tais valores, passamos para a fase de treino do AG. O processo evolutivo do AG está descrito na Seção 3.1.1.4. A Tabela 3.3.2 apresenta as quatro melhores soluções encontradas para cada um dos quatro *folds* de treino usados. A melhor solução

Tabela 3.1. Melhores soluções encontradas pelo AG para cada *fold* de treino

	α	β	γ	δ	ϵ	θ	F_1	Limiar
Fold 1	0,0125	0,0826	0,0469	0,9478	0,2052	0,6584	0,2236	0,02
Fold 2	0,1959	0,1503	0,0089	0,7070	0,3543	0,9176	0,2117	0,03
Fold 3	0,2492	0,0606	0,2305	0,2982	0,4718	0,9687	0,2097	0,06
Fold 4	0,1708	0,1221	0,0218	0,6506	0,4361	0,9053	0,2104	0,03

do problema (pesos das evidências e limiar) corresponde àquela associada ao maior valor gerado pela função objetivo F_1.

Pelos resultados obtidos concluímos que a evidência com predominância de maior peso é a Texto de Âncora (θ) e a Title (δ), isto significa que as informações obtidas a partir destas fontes descrevem melhor o conteúdo de uma página p. Esta conclusão é justificada, pois muitas vezes tanto o conteúdo dos apontadores que referenciam p quanto o título de uma página p são descrições simples e objetivas de seu conteúdo. Percebemos ainda que a evidência TF×IDF (ϵ) é a que possui o segundo maior peso, justificado pois o objetivo desta função é mensurar a importância de t como *tag* de p.

Logo, a Equação 3.1 pode ser reescrita de acordo com os melhores valores encontrados para cada um dos quatro *folds*. Finalizando este experimento obtivemos as Equações (3.3), (3.4), (3.5) e (3.6) como modelo usado para recomendar conjuntos de *tags* para páginas *Web*.

$$\begin{aligned} Imp(t,D) &= [0,0125 \times eH1(t,D)] + [0,0826 \times eH2(t,D)] + [0,0469 \times eBold(t,D)] \\ &\quad + [0,9478 \times eTitle(t,D)] + [0,2052 \times vTF \times IDF(t)] \\ &\quad + [0,6584 \times eAncTex(t,D)] \end{aligned} \tag{3.3}$$

$$\begin{aligned} Imp(t,D) &= [0,1959 \times eH1(t,D)] + [0,1503 \times eH2(t,D)] + [0,0089 \times eBold(t,D)] \\ &\quad + [0,7070 \times eTitle(t,D)] + [0,3543 \times vTF \times IDF(t)] \\ &\quad + [0,9176 \times eAncTex(t,D)] \end{aligned} \tag{3.4}$$

$$\begin{aligned} Imp(t,D) &= [0,2492 \times eH1(t,D)] + [0,0606 \times eH2(t,D)] + [0,2305 \times eBold(t,D)] \\ &\quad + [0,2982 \times eTitle(t,D)] + [0,4718 \times vTF \times IDF(t)] \\ &\quad + [0,9687 \times eAncTex(t,D)] \end{aligned} \tag{3.5}$$

$$
\begin{aligned}
Imp(t,D) &= [0,1708 \times eH1(t,D)] + [0,1221 \times eH2(t,D)] + [0,0218 \times eBold(t,D)] \\
&\quad +[0,6506 \times eTitle(t,D)] + [0,4361 \times vTF \times IDF(t)] \\
&\quad +[0,9053 \times eAncTex(t,D)] \qquad (3.6)
\end{aligned}
$$

3.3.3 Experimento 2: Avaliação dos métodos

O segundo experimento executado teve a finalidade de testar a eficiência dos métodos *agTag* e *agDirTag* descrito nas seções 3.1 e 3.2, respectivamente.

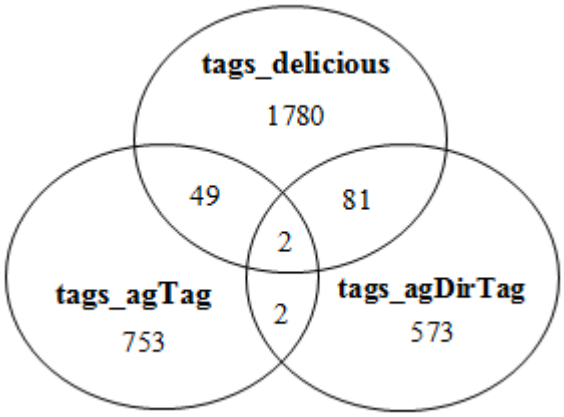
Para avaliar a eficiência de ambos os métodos os conjuntos *tags_ agTag* e *tags_ agDirTag* foram submetidos ao julgamento de avaliadores. Cada avaliador analisou as *tags* recomendadas a dez páginas *Web* distintas, todas provindas dos *folds* de teste. Cada página teve suas *tags* analisadas três vezes por avaliadores diferentes.

Para cada avaliador foi apresentado a página *Web p* para a qual foram sugeridas novas *tags* e o conjunto de *tags* de *p*. O conjunto de *tags* avaliado é formado pelas *tags* obtidas do *site Del.ici.ous* (*tags_ delicious*), pelas *tags* recomendados pelo método *agTag* (*tags_ agTag*) e pelas *tags* recomendados pelo método *agDirTag* (*tags_ agDirTag*). Foi solicitado aos avaliadores que os mesmos lessem e compreendessem o conteúdo das páginas para que, em seguida, assinalassem as *tags* que julgassem ser relevante para a página. Desta forma avaliamos a qualidade do conjunto de *tags* recomendadas pelos métodos e a qualidade do conjunto de *tags* definidos pelo *site Del.ici.ous* para uma página *p*, considerado ideal, por nós, durante a fase de treino do AG.

Neste experimento, foram avaliados os conjuntos *tags_ agTag* e *tags_ agDirTag* das 112 páginas pertencentes a base de referência. Sendo usado para cada um dos quatro *folds* de teste a sua respectiva função $Imp(t, D)$ e seu respectivo valor de limiar, conforme experimento da Seção 3.3.2.

A Figura 3.4 mostra o total de *tags* contidas nos conjuntos *tags_ agTag*, *tags_ agDirTag* e *tags_ delicicous*, bem como as suas interseções. Podemos perceber que os três conjuntos de *tags* são bastante disjuntos. Esta característica se justifica pois a origem das evidências utilizadas para compor cada conjunto é distinta: as *tags* do conjunto *tags_ delicious* refletem opiniões pessoais de usuários acerca de uma página *p*, as *tags* do conjunto *tags_ agTag* tem como base principal o texto de *p*. Já as *tags* do conjunto *tags_ agDirTag* refletem a opinião de um diretório *Web* sobre o conteúdo de *p*. Esta disjunção também significa que os métodos são capazes de sugerir novas *tags* para páginas *Web*: 1328 novas *tags* foram sugeridas, que corresponde a 69,45% do total de *tags* pertencentes ao conjunto *tags_ delicicous*.

Figura 3.4. Quantidade de *tags* sugeridas pelos métodos *agTag*, *agDirTag* e pelo *site Del.ici.ous*.



Sumarizando os dados do julgamento feito pelos avaliadores, podemos obter os seguintes resultados acerca dos conjuntos de *tags* avaliados:

1. Interseção dos conjuntos de *tags*: ($tags\_delicious \cap tags\_agTag$), ($tags\_delicious \cap tags\_agDirTag$) e ($tags\_delicious \cap tags\_agTag \cap tags\_agDirtag$);

2. Porcentagem de $tags \in tags\_agTag$ consideradas boas e que $\notin (tags\_delicious \cap tags\_agTag)$;

3. Porcentagem de $tags \in tags\_agDirTag$ consideradas boas e que $\notin (tags\_delicious \cap tags\_agDirTag)$;

4. Porcentagem de $tags \in tags\_delicious$ consideradas boas;

5. Porcentagem de $tags \in tags\_agTag$ consideradas boas;

6. Porcentagem de $tags \in tags\_agDirTag$ consideradas boas;

7. Acréscimo de *tags* boas provenientes de $tags\_agTag$ caso ($tags\_delicious \cup tags\_agTag$);

8. Acréscimo de *tags* boas provenientes de $tags\_agDirTag$ caso ($tags\_delicious \cup tags\_agDirTag$).

A seguir analisaremos os resultados obtidos a partir dos julgamentos dos avaliadores.

3.3.3.1 Julgamento feito por um avaliador

Neste experimento consideramos que uma *tag* t é boa para uma página p se na opinião de pelo menos um avaliador este fato for verdade.

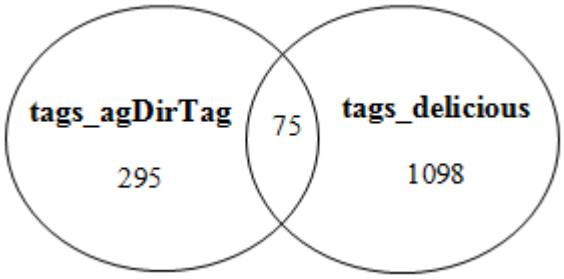
A Figura 3.5 demonstra que o método *agTag* foi capaz de sugerir 362 boas *tags*, das quais, 39 já eram conhecidas e pertenciam ao conjunto *tags_ delicious* e 323 são novas *tags*. Calculando a porcentagem de *tags* consideradas boas pelos avaliadores e que não pertencem ao conjunto *tags_ delicious* obtemos o valor 42,78%. Isso, significa que o método *agTag* foi capaz de sugerir novas e boas *tags* e ainda que, na opinião dos avaliadores, o conjunto *tags_ delicious* não é exaustivo quanto a quantidade de *tags* boas para páginas *Web*. Do total de *tags* pertencentes ao conjunto *tags_ delicious* apenas 61,35% são realmente consideradas boas *tags* pelo julgamento dos avaliadores. Caso houvesse a união dos conjuntos *tags_ delicious* e *tags_ agTag*, o método *agTag* seria capaz de acrescentar cerca de 27% de novas e boas *tags* para o subconjunto de *tags_ delicious*.

Figura 3.5. Quantidade de *tags* pertencentes aos conjuntos *tags_ agTag* e *tags_ delicious* quando pelo menos um usuário as avaliam como boas

Em relação ao método *agDirTag*, a Figura 3.6 demonstra que este foi capaz de sugerir 370 boas *tags*, das quais, 75 já eram conhecidas e pertenciam ao conjunto *tags_ delicious* e 295 são novas *tags*. Calculando a porcentagem de *tags* consideradas boas pelos avaliadores e que não pertencem ao conjunto *tags_ delicious* obtemos o valor 51,30%. Isso, significa que o método *agDirTag* também é capaz de sugerir novas e boas *tags* para páginas *Web*. Do total de *tags* pertencentes ao conjunto *tags_ agDirTag* mais da metade, 56,23%, são realmente consideradas boas *tags* pelo julgamento dos avaliadores. Caso houvesse a união dos conjuntos *tags_ delicious* e *tags_ agDirTag*, o método *agDirTag* seria capaz de acrescentar 25,14% de novas e boas *tags* para *tags_ delicious*.

Uma característica que chama a atenção em relação aos conjuntos é que são bastante disjuntos. Observando a Figura 3.7 vemos que apenas uma *tag* é comum aos três conjuntos, que apenas 39 *tags* são comuns aos conjuntos *tags_ delicious* e *tags_ agTag*, apenas 75 *tags* são comuns aos conjuntos *tags_ delicious* e *tags_ agDirTag* e apenas 3

Figura 3.6. Quantidade de *tags* pertencentes aos conjuntos *tags_ agDirTag* e *tags_ delicious* quando pelo menos um usuário as avaliam como boas



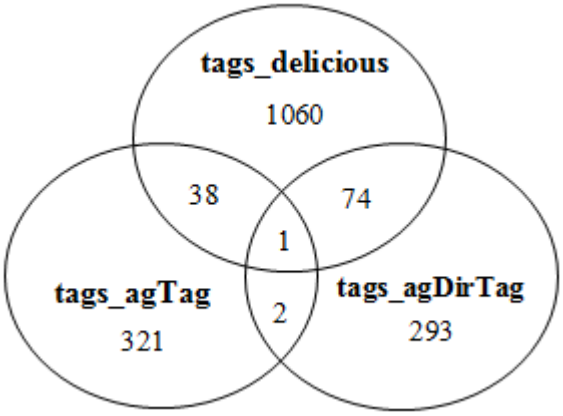
são comuns aos conjuntos *tags_ agTag* e *tags_ agDirTag*. Significando que os métodos são capazes de sugerir novas *tags* para páginas *Web*. Outra característica a se destacar é a baixa taxa de qualidade do conjunto *tags_ delicious* obtidas do *site Del.ici.ous*. Acreditamos que um dos motivos possíveis que justifica este comportamento se deve ao fato de haver *tags* pessoais que tenham sentido pra o usuário que as criou e que não tenham sentido pra outros usuários. Essa questão será estudada mais a fundo em nossos trabalhos futuros. Um outro motivo seria a postagem proposital de *tags* de baixa qualidade, pois encontramos alguns exemplos de *tags* que comprovam tal comportamento: *"*"*, *"-"*, *"educator's_ blogs"*, *1*, *"and"*. Além disso, existem usuários que mesmo sem compreender o assunto da página a ser anotada postam *tags* que muitas vezes não refletem corretamente o conteúdo das páginas, como exemplo: *"all"*,*"4th"*, *"best"*. Ainda citamos como exemplo de *tags* de baixa qualidade as escritas com erro ortográfico. Nestes experimentos observamos também que a porcentagem de *tags* do conjunto *tags_ delicious* consideradas boas assemelha-se a porcentagem de *tags* boas do conjunto *tags_ agTag* e do conjunto *tags_ agDirTag*. Indicando que, em porcentagem, os conjuntos *tags_ agTag* e *tags_ agDirTag* são quase tão bons quando o conjunto *tags_ delicious*.

3.3.3.2 Julgamento feito por dois avaliadores

Neste experimento consideramos que uma *tag* t é boa para uma página p se na opinião de pelo menos dois avaliadores este fato for verdade.

Quando levamos em consideração o consenso de dois usuários a respeito da qualidade das *tags* é visível a queda do número de *tags* consideradas boas. Considerando este cenário, a Figura 3.8 demonstra que o método *agTag* foi capaz de sugerir 215 boas *tags*, sendo 188 boas e novas *tags*. Ao calcularmos a porcentagem de boas e novas *tags* encontramos o valor 24,90%. Ou seja, o método *agTag* foi capaz de sugerir cerca de $\frac{1}{4}$ do total de *tags* já existentes de novas e boas *tags*. Verificamos ainda, que o consenso

Figura 3.7. Quantidade de *tags* pertencentes aos conjuntos *tags_ agDirTag*, *tags_ agDirTag* e *tags_ delicious* quando pelo menos um usuário as avaliam como boa

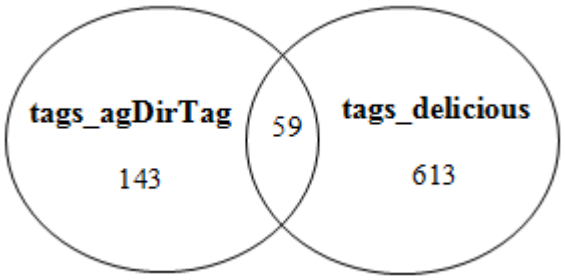


em relação a boa qualidade das *tags* $\in$ *tags_ delicious* diminuiu, apenas 35,15% das *tags* foram consideradas boas pelos avaliadores. Caso houvesse a união dos conjuntos *tags_ delicious* e *tags_ agTag*, o método *agTag* seria capaz de acrescentar 27,9% de novas e boas *tags* para o subconjunto de *tags_ delicious* composto por *tags* avaliadas boas, valor semelhante ao produzido pelo mesmo método considerando o julgamento de um avaliador.

Figura 3.8. Quantidade de *tags* pertencentes aos conjuntos *tags_ agTag* e *tags_ delicious* quando pelo menos dois usuário avaliam uma *tag* como boa

Pela Figura 3.9 percebemos que método *agDirTag* sugeriu 143 boas e novas *tags*, houve interseção de apenas 59 *tags* com o conjunto *tags_ delicious*. Calculando a porcentagem de *tags* consideradas boas pelos avaliadores e que não pertencem ao conjunto *tags_ delicious* obtemos o valor 24,86%. Do total de *tags* pertencentes ao conjunto *tags_ agDirTag* 31% foram consideradas boas *tags* pelos avaliadores. Por outro lado, do total de *tags* pertencentes ao conjunto *tags_ delicious* 35% foram consideradas boas. Caso houvesse a união dos conjuntos *tags_ delicious* e *tags_ agDirTag*, o método *agDirTag* seria capaz de acrescentar 21,27% de novas e boas *tags* para *tags_ delicious*.

Figura 3.9. Quantidade de *tags* pertencentes aos conjuntos *tags_ agDirTag* e *tags_ delicious* quando pelo menos dois usuários avaliam uma *tag* como boa

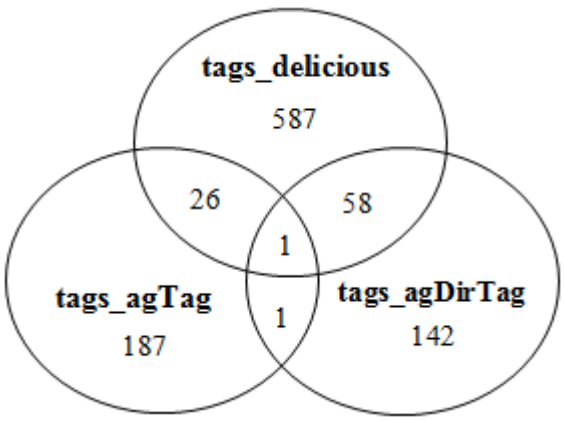


Na Figura 3.10 mostramos o comportamento dos conjunto quando os três são avaliados juntos. Aqui os conjuntos também são bastante disjuntos. Pela a Figura 3.10 vemos que uma *tags* é comum aos três conjuntos, que apenas 23 *tags* são comuns aos conjuntos *tags_ delicious* e *tags_ agTag*, apenas 59 *tags* são comuns aos conjuntos *tags_ delicious* e *tags_ agDirTag* e apenas 2 são comuns aos conjuntos *tags_ agTag* e *tags_ agDirTag*.

Quanto às *tags* pertencentes ao conjunto *tags_ delicious*, observamos que o consenso em relação a sua qualidade é o mesmo encontrado no experimento anterior: baixa. Fato que nos surpreende, pois vários trabalhos adotam o as *tags* do site *Del.ici.ous* como ideais para páginas *Web* [15],[20].

Ainda verificamos que a porcentagem de *tags* do conjunto *tags_ delicious* consideradas boas assemelha-se a porcentagem de *tags* boas do conjunto *tags_ agTag* e a do conjunto *tags_ agDirTag*. Confirmando a indicação inicial que os conjuntos *tags_ agTag* e *tags_ agDirTag* são quase tão bons quando o conjunto *tags_ delicious*.

Figura 3.10. Quantidade de *tags* pertencentes aos conjuntos *tags_ agDirTag*, *tags_ agDirTag* e *tags_ delicious* quando pelo menos dois usuários avaliam uma *tag* como boa



Em um próximo experimentos, usaremos o trabalho de Belem *et al* como *baseline* na avaliação dos métodos de recomendação de *tags* (*agTag* e *agDirTag*). Iremos verificar quão bons são nossos resultados se comparados a este trabalho.

Acreditamos que as *tags* de uma página *Web* possuem vasta aplicabilidade nos sistemas de RI em geral, podendo servir como uma nova fonte de evidência para melhorar o *ranking* sistemas de busca de informações, aperfeiçoar sistemas de classificação e de filtragem de páginas. Ambos os métodos aqui propostos são capazes de recomendar novas e boas *tags* para páginas *Web*. Como próxima etapa deste trabalho vamos estudar o impacto de tais *tags* em um sistema de *ranking* de busca de informações.

Verificamos ainda que o conjunto de *tags* obtidos do *site Del.ici.ous* para uma página p não é um conjunto livre de ruídos, isto é, a qualidade de seus elementos foi posta em dúvida quando analisados pelos avaliadores. Isto significa, que não podemos fixar este conjunto como o ideal para representar as *tags* de páginas *Web*. Além disso, percebemos que várias outras *tags* podem ser acrescentadas a este conjunto, tornando-o mais completo.

Outra característica observada nos métodos propostos é que ambos são capazes de sugerir novas *tags* para páginas *Web*, mesmo que estas possuam pouco conteúdo textual, pois as informação obtidas da concatenação do texto de âncora e da hierarquia do diretório ODP suprem a falta de texto destas últimas.

Capítulo 4

Conclusão

Como visto no Capítulo 1, várias são as pesquisas realizadas na área de Recuperação de Informação voltadas ao uso de *tags*. Elas se tornaram tão populares com a Web 2.0 que seu uso na solução de diversos problemas se tornou comum. Problemas como o de busca, de navegação, de sumarização, de formação de opiniões, de reconhecimento/correspondência podem conter em suas soluções informações provindas de *tags*.

Como primeiro objetivo propusemos dois métodos de anotação automática de páginas *Web*. Os métodos utilizam diversas fontes de evidências textuais, extraídas de uma página p, que combinadas originam listas de termos associados ao assunto de p , e que formam os conjuntos de *tags* de p. A combinação das evidências utilizadas foi feita automaticamente através do uso de algoritmos genéticos, evitando assim as suposições humanas. Avaliando o método verificamos que ambos foram capazes de sugerir novas e boas *tags* para páginas *Web*. Pretendemos, agora, avaliar o impacto da utilização destas *tags* em sistemas de recuperação de informação, mais especificamente explorando seu uso como nova fonte de evidência de relevância em problemas de busca, também como mecanismo para aprimorar *interfaces* em máquinas de busca e como característica em sistemas de classificação de páginas *Web*.

4.1 Próximos Passos

Nossos objetivos a partir de agora são:

1. Mudança no treino do AG: queremos mudar o conjunto de *tags* inicialmente obtido do *site Del.ici.ous* e tomado como conjunto ideal de *tags* para as páginas *Web*, pois experimentos mostraram que nele estão contidas muitas *tags* ruidosas.

O novo conjunto ideal de *tags* para as páginas da base de treino do AG seria composto apenas pelas *tags* consideradas boas no julgamento dos avaliadores, isto é, todas as *tags* julgadas boas dos conjuntos *tags_ delicious*, *tags_ agTag* e *tags_ agDirTag*.

2. Baixar o diretório *ODP*: atualmente as buscas as categorias dos diretório são feitas *on-line*, demandando tempo. Nossa intensão é baixar o conteúdo do diretório, indexá-lo e fazer consultas locais.

3. Mudança na base de referência: os experimentos feitos se restringiram apenas a 112 página, queremos utilizar uma maior de referência maior como a *ClueWeb* [13].

4. Faremos ainda um estudo para caracterizar o uso de *tags* em páginas *Web* (*tags* pessoais, *tags* coletivas, propósito das *tags*, localização, qualidade das *tags*).

   Como resultado dos experimentos realizados obtivemos que as *tags* do *site* *Del.ici.ous* foram consideradas ruins em sua maioria. Pretendemos investigar detalhadamente o que faz uma *tag* ser considerada boa ou ruim. Quais tipos de problemas um conjunto de *tags*, obtidos de *sites* que permitam anotação colaborativa de objetos, pode apresentar. Como é possível diferenciar *tags* boas de ruins dentro destes *sites*? Qual a melhor forma de se representar *tags*, usando termos simples ou compostos? Que características são mais comuns a cada tipo de representação (abrangência das *tags*, grafia)?

5. Avaliar o impacto da utilização de *tags* em sistemas de recuperação de informação. Gupta *et al*, em [9], enumera as utilidade de um conjunto de *tags* em sistemas de RI: busca, classificação, geração de taxonomias, descoberta de interesses sociais comuns, *browsing*, indexação.

   Nosso objetivo em avaliar o impacto da utilização de tais *tags* se restringirá inicialmente a problemas de busca e de classificação de páginas *Web*.

   Inicialmente investigaremos o uso de *tags* como fonte de evidência em problemas de busca na *Web*. Usaremos técnicas de *machine learning* para combinar tal evidência às evidências frequentemente utilizadas na solução deste problema. Nosso objetivo é verificar se os resultados do sistema melhoram ao acrescentarmos tal informação. E consequentemente, sugerir formas adequadas de combinação destas evidências. Bischoff *et al* [5], avaliam se *tags* podem ser usadas na busca de recursos. Para isso, avaliaram a sobreposição de um conjunto de *tags* obtidas de *sites* que permitem anotação colaborativa de objetos com o *log* de consultas da

[13] http://lemurproject.org/clueweb09/

AOL. Porém, acreditamos que o simples fato de uma *tag* estar presente em uma consulta não necessariamente indica que o uso de *tags* como fonte de evidencia traria ganhos de qualidade em sistemas de RI. Em seguida, faremos o mesmo experimento, porém aplicado a problemas de classificação de páginas *Web*.

Referências Bibliográficas

[1] Open directory project, 2002. Disponivel em: http://www.dmoz.org/. Data acesso: Novembro 2011.

[2] Malik Agyemang, Ken Barker, and Rada S. Alhajj. Mining web content outliers using structure oriented weighting techniques and n-grams. In *Proceedings of the 2005 ACM symposium on Applied computing*, pages 482–487, New York, NY, USA, 2005.

[3] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval: the concepts and technology behind search.* Pearson Education, England, 2011.

[4] Fabiano Belém, Eder Martins, Tatiana Pontes, Jussara Almeida, and Marcos Gonçalves. Associative tag recommendation exploiting multiple textual features. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information*, SIGIR '11, pages 1033–1042, 2011.

[5] Kerstin Bischoff, Claudiu S. Firan, Wolfgang Nejdl, and Raluca Paiu. Can all tags be used for search? In *Proceeding of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 193–202. ACM, 2008.

[6] Hendrik Blockeel and Jan Struyf. Efficient algorithms for decision tree cross-validation. In *Journal of Machine Learning Research*, pages 621–650, 2002.

[7] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.*, 30:107–117, April 1998.

[8] William Cavnar, , William B. Cavnar, and John M. Trenkle. N-gram-based text categorization. In *In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, 1994.

[9] Manish Gupta, Rui Li, Zhijun Yin, and Jiawei Han. Survey on social tagging techniques. *SIGKDD Explor. Newsl.*, 12:58–72, November 2010.

[10] Jong Wook Kim and K. Selcuk Candan. Cp/cv: concept similarity mining without frequency information from domain describing taxonomies. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, CIKM '06. ACM, 2006.

[11] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46:604–632, September 1999.

[12] Byron Y-L Kuo, Thomas Hentrich, Benjamin M . Good, and Mark D. Wilkinson. Tag clouds for summarizing web search results. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 1203–1204. ACM, 2007.

[13] R Linden. *Algoritmos GenÃ©ticos: Uma importante ferramenta da InteligÃªncia Computacional*. McGraw-Hill Science, Rio de Janeiro, Brasil, 2006.

[14] Marek Lipczak and Evangelos Milios. Learning in efficient tag recommendation. In *Proceedings of the fourth ACM conference on Recommender systems*, RecSys '10, pages 167–174. ACM, 2010.

[15] Guilherme Vale Menezes, Jussara M. Almeida, Fabiano Belém, Marcos André Gonçalves, Anísio Lacerda, Edleno Silva De Moura, Gisele L. Pappa, Adriano Veloso, and Nivio Ziviani. Demand-driven tag recommendation. In *Proceedings of the 2010 European conference on Machine learning and knowledge discovery in databases: Part II*, ECML PKDD'10. Springer-Verlag, 2010.

[16] Davood Rafiei and Alberto O. Mendelzon. What is this page known for? computing web page reputations. In *Proceedings of the 9th international World Wide Web conference on Computer networks : the international journal of computer and telecommunications netowrking*, pages 823–835. North-Holland Publishing Co., 2000.

[17] A. W. Rivadeneira, Daniel M. Gruen, Michael J. Muller, and David R. Millen. Getting our head in the clouds: toward evaluation studies of tagclouds. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '07, pages 995–998. ACM, 2007.

[18] Börkur Sigurbjörnsson and Roelof van Zwol. Flickr tag recommendation based on collective knowledge. In *Proceeding of the 17th international conference on World Wide Web*, WWW '08, 2008.

[19] Petros Venetis, Georgia Koutrika, and Hector Garcia-Molina. On the selection of tags for tag clouds. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 835–844. ACM, 2011.

[20] Robert Wetzker, Carsten Zimmermann, Christian Bauckhage, and Sahin Albayrak. I tag, you tag: translating tags for advanced user models. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 71–80. ACM, 2010.