COPPE/UFRJ

REPRESENTAÇÃO DE DOCUMENTOS ATRAVÉS DE NUVENS DE TERMOS

Fernando Fernandes Morgado

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Sistemas e Computação.

Orientador: Geraldo Bonorino Xexéo

Rio de Janeiro

Setembro de 2010

REPRESENTAÇÃO DE DOCUMENTOS ATRAVÉS DE NUVENS DE TERMOS

Fernando Fernandes Morgado

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Examinada por:

_____

Prof. Geraldo Bonorino Xexéo, D. Sc.

_____

Prof. Jano Moreira de Souza, D.Sc.

_____

Prof. Adriana Santarosa Vivacqua, D.Sc.

RIO DE JANEIRO, RJ - BRASIL
SETEMBRO DE 2010

Morgado, Fernando Fernandes

Representação de Documentos Através de Nuvens de Termos/ Fernando Fernandes Morgado. – Rio de Janeiro: UFRJ/COPPE, 2010.

XIV, 119 p.: il.; 29,7 cm.

Orientador: Geraldo Bonorino Xexéo

Dissertação (mestrado) – UFRJ/ COPPE/ Programa de Engenharia de Sistemas e Computação, 2010.

Referencias Bibliográficas: p. 90-93.

1. Representação de documentos. 2. Seleção de termos. 3. Tag Clouds. I. Xexéo, Geraldo Bonorino II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Titulo.

Àquela que sempre me apoiou e incentivou,
minha mãe Maria Aparecida.
E ao meu pai Abílio (*in memoriam*), por
tudo o que fez por mim e me ensinou.

Agradecimentos

Primeiramente gostaria de agradecer a Deus por me ter concedido a oportunidade de concluir este trabalho e por tudo o que Ele representa em minha vida e fez por mim.

Agradeço também a minha família, principalmente a minha mãe Maria Aparecida e minhas irmãs, por todo o suporte e apoio que me deram durante esta etapa para que eu pudesse concluí-la. Agradeço também ao meu saudoso pai Abílio, pois, ele me acompanhou no início desta caminhada, mas infelizmente não está mais aqui para ver o término.

Agradeço aos meus amigos de mestrado tantos os alunos de Banco de Dados quanto os dos demais cursos, especialmente aos que entraram junto comigo e me ajudaram em diversos momentos: Cimar, Clarissa, Edno, Ester, Fred, Luciano, Marcelino, Vanessa e Viviane.

Dedico um agradecimento especial a doutoranda Patrícia Fiuza que me ajudou em diversos momentos, seja dando idéias, explicando dúvidas, sugerindo novas abordagens e pela paciência por aturar a mim e as longas conversas trocadas através de emails.

Agradeço aos amigos Marcello e Tiago que sempre demonstraram interesse em acompanhar o andamento desta minha jornada. Também aos meus amigos da graduação do curso de Ciência da Computação da UFRJ, da COPPETEC e do Serpro que me acompanharam nesta reta final.

Especialmente, agradeço aos amigos que puderam me ajudar na conclusão deste trabalho, participando da avaliação realizada, dentre eles: Alan, Anderson, Bruno, Cimar, Eduardo, Ester, Fellipe, Julliano, Leonardo, Luciano, Marcelino, Patrícia, Rodrigo, Thatiana, Vanessa e Viviane.

Também dedico um especial agradecimento ao meu orientador Geraldo Xexéo, pela orientação neste trabalho e pelas inúmeras contribuições fornecidas, as quais permitiram a conclusão deste trabalho. Agradeço também ao professor Jano por ter aceitado participar da banca assim como a professora Adriana.

Agradeço também a todos que de forma direta ou indireta ajudaram na conclusão deste trabalho, principalmente aos funcionários das secretarias do Programa

de Engenharia de Sistemas e Computação da COPPE e da linha de Banco de Dados, entre os quais, Solange, Patrícia, Ana Paula e os demais.

Por fim, agradeço à Fundação CAPES pela bolsa de mestrado concedida durante esta etapa.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

REPRESENTAÇÃO DE DOCUMENTOS ATRAVÉS DE NUVENS DE TERMOS

Fernando Fernandes Morgado

Setembro/2010

Orientador: Geraldo Bonorino Xexéo

Programa: Engenharia de Sistemas e Computação

Com o crescimento das ferramentas ligadas a web social a ação de tagging vem sendo cada vez mais utilizada. Com o crescimento desta atividade surgiu a necessidade de criar uma nova forma de apresentar estas tags aos usuários. Desta necessidade surgiram as Tag Clouds.

Ao analisarmos o cenário atual das ferramentas de busca e recuperação de informação percebemos que elas não permitem que os usuários tenham uma visão global da lista de documentos (resultados) retornados pela consulta.

Portanto, neste trabalho é proposta a utilização de Tag Clouds em conjunto com uma ferramenta de busca e recuperação de informação para possibilitar duas funcionalidades aos usuários. A primeira que eles consigam visualizar um resumo dos resultados retornados na consulta. A segunda que consigam distinguir características e conceitos que diferenciam um específico documento em relação aos demais documentos retornados no resultado.

Para possibilitar a criação destas Tag Clouds, neste trabalho também será apresentado um modelo formal que permita a criação destas Tag Clouds propostas.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

DOCUMENTS REPRESENTATION THROUGH TAG CLOUDS

Fernando Fernandes Morgado

September/2010

Advisor: Geraldo Bonorino Xexéo

Department: Systems and Computer Engineering

With the growth number of social web tools, the action of tagging is being used increasingly. With this activity came the need to create a new way of presenting these tags to users. From this need came the Tag Clouds.

Analyzing the current scenario of the search tools and information retrieval, we realize that they do not allow users to have an overview of the list of documents (results) returned by the query.

Therefore, this paper proposes the use Tag Clouds combined with a information retrieval tool, to enable two features to users. The first feature to allow that users can view a summary of results returned in the query. The second feature, that enables distinguish features and concepts that differentiate a specific document in relation to other documents returned in the result.

To enable the creation of these Tag Clouds in this paper, we presented a formal model that allows the creation of these proposals Tag Clouds.

Sumário

Lista de Figuras

Lista de Tabelas

1. Introdução

1.1 Motivação

Com o crescimento das ferramentas ligadas a *web* social, uma ação vem sendo cada vez mais utilizada: a ação de *tagging*. A ação de *tagear*, ou *tagging*, pode ser definida como a atribuição de *tags* a objetos, a fim de descrever as propriedades ou atributos deste objeto. Estes objetos podem ser documentos, músicas, fotos, vídeos entre outros (Halvey, 2007).

Segundo (Penev, 2008), uma *tag* é qualquer idéia simples descrevendo um objeto. Ele ainda destaca que uma combinação destas *tags* irá descrever o objeto em maiores detalhes.

Em decorrência do incentivo ao ato de *tagging*, surgiu a necessidade da criação de métodos para exibir estas *tags* aos usuários. Uma técnica adotada, em diversos sites, para a realização desta tarefa é a utilização de *Tag Clouds*. *Tag Clouds* são representações visuais de um conjunto de palavras, tipicamente um conjunto de *tags*, no qual atributos do texto como tamanho, peso ou cor podem ser utilizados para representar características dos termos associados (Halvey, 2007).

Por outro lado, ao analisarmos o cenário atual das ferramentas de busca e recuperação de informação, observamos que a maioria destas ferramentas possui a característica em comum de apresentar os documentos retornados na busca como uma lista ordenada baseado em um valor numérico de relevância. Geralmente, os documentos desta lista serão divididos em várias páginas. Esta forma de apresentação dos resultados impossibilita o usuário de ter uma visão global da lista completa de documentos retornados pela ferramenta. Isto, provavelmente, irá forçar o usuário a navegar por diversas páginas em busca de documentos que sejam relevantes para ele.

Além disto, em (Silverstein *et al.*, 1999) apresenta uma análise estatística que afirma que em 85% das situaçãoes apenas a primeira página de resultados é vista pelo usuário.

Portanto, podemos considerar relevante que, nessa primeira página, o usuário tenha de alguma forma um resumo dos resultados retornados na consulta. E da mesma forma, também podemos avaliar como útil que o usuário consiga distinguir as características e conceitos que diferenciam um específico documento em relação aos demais documentos retornados no resultado.

Uma forma de se viabilizar estes dois requisitos é a aplicação de tag clouds. A idéia consiste essencialmente em adicionar dois tipos diferentes de tag clouds a uma típica ferramenta de busca e recuperação de informação. Cada uma das tag clouds atenderia uma finalidade específica. São elas:

1. Uma primeira *tag cloud* para resumir os assuntos e conceitos abordados nos documentos retornados na busca. Chamaremos esta *tag cloud* de **"*Tag Cloud* de Resumo do Conjunto"**.
2. Uma segunda *tag cloud* para permitir a visualização das características que distinguem um determinado documento em relação aos demais documentos retornados na busca. Chamaremos esta *tag cloud* de ***"Tag Cloud* Diferencial"**. A ideia deste teipo de tag cloud pode ser encontrada também em (Xexéo *et al.*, 2009b).

Para possibilitar a criação destes dois tipos de *tag clouds*, este trabalho estabelece, primeiramente, um modelo formal que permita a criação das mesmas. Este modelo foi apresentado parcialmente, durante sua elaboração em (Xexéo *et al.*, 2009a). O objetivo é fornecer tal modelo de forma que possa ser utilizado na criação de outras ferramentas baseadas em tag clouds. Tal modelo é baseado no processo de criação de *tag clouds* descrito em (Marinchev, 2006). Segundo (Marinchev, 2006), uma *tag cloud*

é uma representação visual de um campo semântico de um objeto. E um campo semântico é definido como: “o conjunto de conceitos conectados a um objeto foco, de tal forma que é independente das pessoas que atribuíram as *tags* (tageadores originais) e possibilita que outras pessoas tenham o entendimento destes conceitos.”.

Seguindo esta definição, podemos entender que, para a construção das *tag clouds*, é necessário que se capture o campo semântico do objeto ao qual a *tag cloud* está associada. Na abordagem proposta neste trabalho, um conjunto destes objetos é representado pelos documentos retornados como resultado da consulta e o campo semântico de cada documento é o conjunto de palavras que melhor descreve os conceitos ou assuntos contidos neste documento.

Portanto, uma das dificuldades encontradas será conseguir fazer esta identificação de quais conceitos estão associados aos documentos. Para isso, vamos utilizar da técnica de *Latent Semantic Analysis (LSA)*, além de outras abordagens que serão apresentadas no decorrer deste trabalho.

1.2 Trabalhos Relacionados

[C1] Comentário: Vai manter esse parágrafo sobre tópicos?

Existem alguns trabalhos anteriores sobre a geração automática de *tag clouds* e a sua utilização em tarefas de busca e recuperação de informação.

Primeiramente, são apresentados alguns conceitos e aspectos relacionados à área de busca e recuperação de informação e ferramentas de busca.

Gerard Salton, um dos precursores da área de I*nformation Retrieval (IR)*, em português *Busca e recuperação de Informação (BRI)*, e um dos líderes desde os anos 60 até os anos 90, propôs em (Salton, 1968) a seguinte definição: *“ Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information.”*

Em (Anderson, 1990) foi apresentada uma analogia entre *information retrieval* e processos da memória semântica humana. Um meio de expressar seu ponto comum é pensar em uma pessoa que quer fazer uma busca e tem em mente algum significado, o qual ela irá expressar em palavras, e o sistema irá tentar encontrar um texto que possua o mesmo significado.

O sucesso desta busca, então, irá depender do sistema de representação da consulta e do significado do texto de um modo que refletirá corretamente suas similaridades para os humanos. A indexação baseada em semântica latente faz isso melhor do que sistemas que dependem da comparação literal entre termos das consultas e dos documentos. Esta superioridade pode ser atribuída à capacidade de combinar corretamente as consultas aos documentos de mesmo significado mesmo quando a consulta e os documentos utilizam palavras diferentes.

Inicialmente, o foco da área de IR era o trabalho sobre documentos textuais. Com a evolução desta área, atualmente aplicações de IR envolvem documentos multimídia com estrutura, conteúdo texto com significado, e outras mídias. Estas outras mídias incluem, fotos, vídeos e áudios.

Entre as aplicações de IR podemos destacar: Web Search, Vertical Search, Enterprise Search, Desktop Search, Peer-to-peer Search, Ad Hoc Search, Classificação e Questions Answering (Croft *et al.*, 2009).

Ainda segundo (Croft *et al.*, 2009) as ferramentas de busca devem suportar dois processos principais, são eles: processo de indexação e processo de consulta.

Em (Kuo *et al.*, 2007), é apresentada uma ferramenta denominada PubCloud, a qual é baseada no uso de uma *tag cloud* para resumir resultados retornados por uma consulta, e também que permite a navegação a partir da *tag cloud* para os resultados obtidos.

Watters em (Watters, 2009) define uma ferramenta chamada CloudMine, a qual tem por objetivo ser uma ferramenta que categoriza, resume e exibe os mais importantes termos de um documento como *text clouds*. Esta visualização auxilia aos usuários na tarefa de avaliação de documentos, enquanto eles realizam uma atividade de recuperação de informação. Uma *text cloud* pode ser considerada uma *tag cloud* aprimorada, isso porque, ela não é composta somente por *tags* de um único termo, mas também por sentenças de mais de um termo. Além disso, seu objetivo principal é possibilitar um maior entendimento do documento ao qual ela está ligada (Lamantia, 2007). Vale destacar que no modelo que apresentaremos neste trabalho será possível a criação de *text clouds* de forma equivalente a criação de *tag clouds*.

Já em outros artigos como (Song *et al.*, 2008) e (Heymann *et al.*, 2008) são apresentadas técnicas de atribuição de *tags* com base em aprendizagem a partir da existência de documentos que já tenham sido *tageados* anteriormente.

Apesar destas abordagens já existentes, a proposta neste trabalho é mais geral no sentido que é criado um modelo formal no qual, diferentes aplicações que utilizem *tag clouds* poderão utiliza-lo para serem desenvolvidas. Este modelo contribui ao ponto que estabelece a compreensão dos requisitos necessários para a construção de um sistema baseado em *tag clouds*.

1.3 Objetivo

Este trabalho possui dois objetivos principais. Um deles propõe uma nova alternativa para a utilização das *tag clouds*. A proposta tem por base utilizar as *tag clouds* em conjunto com uma típica ferramenta de busca e recuperação de informação com o objetivo de auxiliar os usuários desta ferramenta na execução da tarefa de recuperar dentre os documentos retornados pela busca os que sejam importantes para ele.

E o outro objetivo deste trabalho irá fornecer o suporte para a construção destas *tag clouds* que foram sugeridas. Uma vez que nosso segundo objetivo é criar, desde o início, um modelo formal para a criação de *tag clouds*.

Para construir este modelo formal de criação de *tag clouds* será utilizado como base o trabalho apresentado por (Marinchev, 2006).

Também serão apresentados os resultados obtidos após a avaliação tanto da qualidade das *tag clouds* geradas quanto da utilização das *tag clouds* sugeridas em conjunto com uma ferramenta de busca e recuperação de informação.

1.4 Organização

Além deste capítulo de introdução, este trabalho apresenta mais 6 capítulos que estão organizados da seguinte forma: o capítulo 2 apresenta os conceitos relacionados à teoria de LSA e como ela pode ser aplicada para descobrir e inferir relação entre os termos nos documentos.

O capítulo 3 apresenta as definições e teorias relacionadas às *tag clouds*, como por exemplo, suas propriedades e possíveis aplicações.

O capítulo 4 apresenta o modelo formal que é proposto para a criação das *tag clouds*.

No capítulo 5 são apresentados os métodos e implementação que foram utilizadas para gerar as *tag clouds* propostas neste trabalho. Também é apresentado um protótipo para a ferramenta de busca e recuperação de informação sugerida.

O capítulo 6 apresenta as avaliações realizadas neste trabalho para medir a qualidade das tag clouds geradas tanto numericamente como subjetivamente, além de comparar com outras implementações disponíveis na web.

E finalmente no capítulo 7 são apresentadas algumas conclusões obtidas e no capítulo 8 a lista de referências bibliográficas utilizada.

2. LSA

Um dos tipos de *tag clouds* proposto neste trabalho, a *tag cloud* de resumo do conjunto, tem por objetivo identificar os assuntos presentes numa coleção de documentos.

Para implementar esta tarefa foi utilizado a teoria de *Latent Semantic Analysis (LSA*). Por isso, nesta seção serão apresentados os conceitos ao redor de LSA, que possibilitou a identificação dos assuntos presentes nos documentos para os quais deveriam ser geradas as *tag clouds* de resumo do conjunto.

2.1 Definições

Análise de Semântica Latente, do inglês *Latent Semantic Analysis (LSA)* (Deerwester *et al.*, 1990), é uma abordagem que combina agrupamento de termos e de documentos. LSA utiliza a matriz termo-documento da representação vetorial como entrada e sobre ela aplica a técnica de redução de dimensão baseando-se na decomposição em valores singulares, *Singular Values Decomposition (SVD)*. Nela, documentos e palavras são mapeados em uma representação do espaço de semântica latente, que é baseada em tópicos ao invés de cada palavra individualmente e por isso, o espaço de representação será muito menor do que o original.

Esta técnica na qual o LSA é baseado é uma técnica de decomposição matemática de matrizes aplicada ao corpo do texto e que tem por objetivo aproximar a experiência que as pessoas têm ao analisar um texto em sua língua.

LSA é uma técnica matemática e/ou estatística completamente automática para extrair e inferir relações de contexto através das palavras usadas em frases e sentenças. LSA não é uma metodologia de processamento de linguagem natural ou um programa de inteligência artificial. Não faz uso de dicionários construídos, bases de

conhecimento, redes semânticas, gramática, *parser* sintático ou morfológico e requer como entrada apenas o texto puro, dividido em palavras e passagens significativas ou apenas como sentenças e parágrafos (Landauer *et al.*, 1998).

A representação de significado de palavras e sentenças que pode ser obtido pelo *LSA* têm sido capaz de simular uma variedade de habilidades que as pessoas possuem como, por exemplo, reconhecimento de vocabulário, categorização de palavras, compreensão de discurso, etc (Landauer *et al.*, 1998).

Para realizar esta representação, o LSA pode produzir medidas entre relações de palavra-palavra, palavra-sentença e sentença-sentença que são bem correlacionadas com os atos que as pessoas praticam quando realizam tarefas de associação ou similaridade semântica.

Ou seja, através do LSA é possível aproximar os julgamentos humanos de similaridade entre palavras e objetivamente prever a similaridade baseada em palavras entre as sentenças, aspecto que é bastante valorizado no estudo de processamento de discurso.

É importante ressaltar que a similaridade estimada derivada do LSA não é uma simples contagem de freqüência, co-ocorrência ou correlação, ela depende de uma poderosa análise matemática. Tal análise é capaz de inferir corretamente muitas relações, daí a expressão “Semântica Latente”, e consequentemente fornece uma melhor simulação dos julgamentos humanos que são baseados em significados.

Entre as limitações do LSA existe o fato de não se usar o ordenamento das palavras, ou seja, de relações sintáticas ou lógicas, ou de morfologia. Apesar disto, ele consegue extrair corretamente os significados de palavras e passagens mesmo sem essas características (Landauer *et al.*, 1998).

No primeiro passo o LSA representa a coleção de documentos através de uma matriz termos x documentos, onde cada célula contém a freqüência que a respectiva palavra da linha aparece no documento referente à coluna.

Em seguida o LSA aplica a decomposição de valor singular, do inglês ***Singular Value Decomposition (SVD)***, a esta matriz. No SVD, uma matriz retangular é decomposta em um produto de três outras matrizes. Uma das matrizes componentes descreve as linhas originais como vetores derivados de um produto por um fator, outra matriz descreve o mesmo só que para as colunas. A terceira é uma matriz diagonal contendo os valores de escala de tal forma que quando as três matrizes são multiplicadas, a matriz original é reconstruída.

A figura 2.1 exibe o processo de decomposição associado à geração do espaço semântico latente.

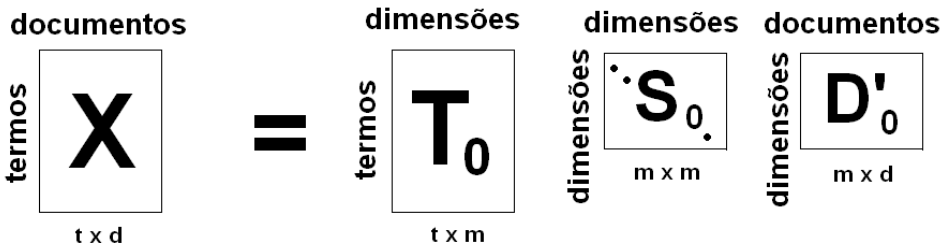


Figura 2.1: Processo de Análise de Semântica Latente

A matriz X, t (termos) * d (documentos) pode ser reescrita pelo produto $T_0 * S_0 * D'_0$.

S_0 é uma matriz diagonal que ordena seus valores de cima para baixo. A matriz X também pode ser reconstruída baseada em número de dimensões menor que o número de dimensões original. Por exemplo, para reconstruir a matriz X baseando-se apenas duas dimensões, basta manter intactos os dois primeiros elementos da matriz S_0 e alterar todos os demais para zero.

Em (Landauer *et al.*, 1998) é apresentado um exemplo que demonstra como esta técnica funciona. A seguir, será apresentado este exemplo.

O exemplo usa como textos nove títulos de documentos técnicos, cinco sobre interação humano computador e quatro sobre teoria matemática de grafos. Tópicos que são conceitualmente distintos.

Desta forma, a matriz termo x documento que será gerada terá nove colunas e 12 linhas, pois, cada linha irá corresponder a uma palavra usada em pelo menos dois títulos. Os títulos com os 12 termos que serão utilizados, destacados em itálico são apresentados na figura 2.2.

ihc1:	*Human* machine *interface* for ABC *computer* applications
ihc2:	A *survey* of *user* opinion of *computer system response time*
ihc3:	The EPS *user interface* management *system*
ihc4:	*System* and *human system* engineering testing of EPS
ihc5:	Relation of *user* perceived *response time* to error measurement
g1:	The generation of random, binary, ordered *trees*
g2:	The intersection *graph* of paths in *trees*
g3:	*Graph minors* IV: Widths of *trees* and well-quasi-ordering
g4:	*Graph minors*: A *survey*

Figura 2.2: Exemplos de texto: Título de documentos técnicos

A matriz termo x documento gerada é apresentada na figura 2.3.

	{X} =								
	ihc1	**ihc2**	**ihc3**	**ihc4**	**ihc5**	**g1**	**g2**	**g3**	**g4**
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

Figura 2.3: Matriz termo x documento gerada

Após a realização da decomposição de valor singular, a matriz {X} é decomposta no seguinte produto de matrizes: {X} = {T} {S} {D}'.

Estas matrizes são exibidas nas figuras 2.4, 2.5 e 2.6. A partir da multiplicação destas matrizes consegue-se obter a matriz {X} novamente.

{T}=								
0.22	-0.11	0.29	-0.41	-0.11	-0.34	0.52	-0.06	-0.41
0.20	-0.07	0.14	-0.55	0.28	0.50	-0.07	-0.01	-0.11
0.24	0.04	-0.16	-0.59	-0.11	-0.25	-0.30	0.06	0.49
0.40	0.06	-0.34	0.10	0.33	0.38	0.00	0.00	0.01
0.64	-0.17	0.36	0.33	-0.16	-0.21	-0.17	0.03	0.27
0.27	0.11	-0.43	0.07	0.08	-0.17	0.28	-0.02	-0.05
0.27	0.11	-0.43	0.07	0.08	-0.17	0.28	-0.02	-0.05
0.30	-0.14	0.33	0.19	0.11	0.27	0.03	-0.02	-0.17
0.21	0.27	-0.18	-0.03	-0.54	0.08	-0.47	-0.04	-0.58
0.01	0.49	0.23	0.03	0.59	-0.39	-0.29	0.25	-0.23
0.04	0.62	0.22	0.00	-0.07	0.11	0.16	-0.68	0.23
0.03	0.45	0.14	-0.01	-0.30	0.28	0.34	0.68	0.18

Figura 2.4: Matriz {W}

{S} =								
3.34	0	0	0	0	0	0	0	0
0	2.54	0	0	0	0	0	0	0
0	0	2.35	0	0	0	0	0	0
0	0	0	1.64	0	0	0	0	0
0	0	0	0	1.50	0	0	0	0
0	0	0	0	0	1.31	0	0	0
0	0	0	0	0	0	0.85	0	0
0	0	0	0	0	0	0	0.56	0
0	0	0	0	0	0	0	0	0.36

Figura 2.5: Matriz {S}

{D}' =								
0.20	0.61	0.46	0.54	0.28	0.00	0.01	0.02	0.08
-0.06	0.17	-0.13	-0.23	0.11	0.19	0.44	0.62	0.53
0.11	-0.50	0.21	0.57	-0.51	0.10	0.19	0.25	0.08
-0.95	-0.03	0.04	0.27	0.15	0.02	0.02	0.01	-0.03
0.05	-0.21	0.38	-0.21	0.33	0.39	0.35	0.15	-0.60
-0.08	-0.26	0.72	-0.37	0.03	-0.30	-0.21	0.00	0.36
0.18	-0.43	-0.24	0.26	0.67	-0.34	-0.15	0.25	0.04
-0.01	0.05	0.01	-0.02	-0.06	0.45	-0.76	0.45	-0.07
-0.06	0.24	0.02	-0.08	-0.26	-0.62	0.02	0.52	-0.45

Figura 2.6: Matriz {P}

A reconstrução da matriz original pode ser feita utilizando-se menos dimensões do que existiam na matriz original. Por exemplo, no caso mostrado é possível utilizar apenas duas dimensões para reconstruir a matriz original. Neste caso, devem ser utilizados apenas os dois primeiros elementos, colunas, das três matrizes mostradas nas

três figuras anteriores.. Isto é equivalente a atribuir 0 para todas as colunas da matriz {S}, exceto para os elementos das duas primeiras colunas. A matriz reconstruída {X'} é mostrada na figura 2.7.

{X'} =									
	ihc1	ihc2	ihc3	ihc4	ihc5	g1	g2	g3	g4
human	0.16	0.40	0.38	0.47	0.18	-0.05	-0.12	-0.16	-0.09
interface	0.14	0.37	0.33	0.40	0.16	-0.03	-0.07	-0.10	-0.04
computer	0.15	0.51	0.36	0.41	0.24	0.02	0.06	0.09	0.12
user	0.26	0.84	0.61	0.70	0.39	0.03	0.08	0.12	0.19
system	0.45	1.23	1.05	1.27	0.56	-0.07	-0.15	-0.21	-0.05
response	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
time	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
EPS	0.22	0.55	0.51	0.63	0.24	-0.07	-0.14	-0.20	-0.11
survey	0.10	0.53	0.23	0.21	0.27	0.14	0.31	0.44	0.42
trees	-0.06	0.23	-0.14	-0.27	0.14	0.24	0.55	0.77	0.66
graph	-0.06	0.34	-0.15	-0.30	0.20	0.31	0.69	0.98	0.85
minors	-0.04	0.25	-0.10	-0.21	0.15	0.22	0.50	0.71	0.62

Figura 2.7: Matriz reconstruída {X'}

Cada valor da matriz reconstruída foi resultado de uma combinação linear de valores das duas dimensões que foram preservadas durante a reconstrução. E por sua vez, estes valores foram calculados como combinação linear dos valores das matrizes originais. Portanto, mudando um valor de qualquer das células da matriz original, os valores na matriz reconstruída também podem ser diferentes. E esta é a forma matemática pela qual o LSA realiza inferência ou indução.

A partir da matriz reconstruída são calculadas as correlações entre os termos. Por exemplo, pode ser utilizada a correlação de Pearson (Rodgers, 1988) para isto. Estas correlações revelam a informação semântica latente na coleção e pode descobrir palavras que, mesmo não estando presentes em determinados documentos, podem estar relacionadas a eles. De acordo com (Griffiths, 2007), estes agrupamentos de termos relacionados podem ser considerados como conceitos ou tópicos.

Uma informação que o LSA utiliza para inferir as relações entre as palavras e as sentenças é a informação sobre as sentenças onde palavras particulares não estão presentes.

Apresentando como uma recente e importante abordagem de modelagem de tópicos, LSA vem sendo utilizada e citada frequentemente em trabalhos de diversas áreas e tem inspirado novas direções de pesquisa. Sua aplicação mais conhecida é a indexação por semântica latente, em ingês *Latent Semantic Indexing (LSI)* da área de recuperação de informação e proposta em (Deerwester *et al.*, 1990) e (Dumais, 1995).

Apesar da obtenção de resultados positivos, ela apresenta alguns problemas principalmente no que diz respeito aos seus fundamentos estatísticos insatisfatórios e a complexidade computacional envolvida no processo (Wei, 2007).

2.2 Considerações Finais

Neste capítulo foi apresentado os conceitos relacionados à técnica de LSA. Foi abordado o aspecto matemático relacionado ao tema bem como as razões para a utilização desta técnica neste trabalho.

3. Tag Clouds

Neste capítulo serão apresentadas algumas definições sobre *tag clouds*, bem como destacar as propriedades relacionadas a elas. Também vamos abordar algumas aplicações suportadas por elas e destacar seus pontos positivos e negativos.

3.1 Definição

Nesta seção são fornecidas algumas definições para *tag clouds* e alguns conceitos relacionados a elas, com o objetivo de promover um entendimento uniforme que irá permitir que posteriormente se possa construir um *framework* formal para a construção delas. Também serão discutidas algumas questões relacionando a diferença entre as *tag clouds* geradas por seres humanos e computadores.

Neste trabalho é adotada a definição proposta por Rivadeneira (Rivadeneira *et al.*, 2007): "*tag clouds* são representações visuais de um conjunto de palavras, tipicamente um conjunto de *tags* selecionadas por algum método racional, na qual atributos do texto como tamanho, estilo, ou cor são usados para representar características dos termos associados." Além disso, uma *tag cloud* possui outro componente, uma referência, por exemplo, um título ou um rótulo que indica a qual objeto esta *tag cloud* está relacionada. Não há a necessidade de o objeto ser concreto ou acessível através de uma URL, tal qual um documento, música ou foto. Por exemplo, uma *tag cloud* pode se referir a um evento que ocorreu ou irá ocorrer, como um show de rock.

Uma ***tag*** pode ser definida como qualquer rótulo ou símbolo anexado a um objeto, como um documento, uma imagem, foto, música, etc. Usualmente, estes rótulos são pequenos e na maioria das vezes são formados por uma única palavra (Watters, 2009). Cada *tag* geralmente procura representar um conceito relacionado ao objeto o

qual ela está ligada. Conceitos podem incluir idéias de origem, propósito, descrição entre outros.

Marinchev em (Marinchev, 2006) identificou o conjunto de conceitos abstratos relacionados a um objeto, e representados em uma *tag cloud*, como um **campo semântico**. Um campo semântico "é o conjunto de conceitos conectados a um objeto foco, de tal forma que é independente das pessoas que atribuíram as *tags* (tageadores originais) e é possível que outras pessoas tenham o entendimento destes conceitos." (Marinchev, 2006). Como resultado, uma *tag cloud* é uma representação visual de um campo semântico de um objeto.

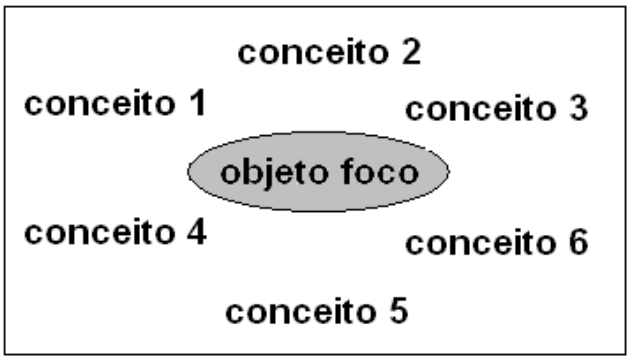


Figura 3.1: Campo Semântico

O processo completo de criação de uma *tag cloud* pode ser resumido em três etapas iniciais (Marinchev, 2006).

1. Compreensão do objeto foco e dos conceitos que podem ser aplicados a ele.

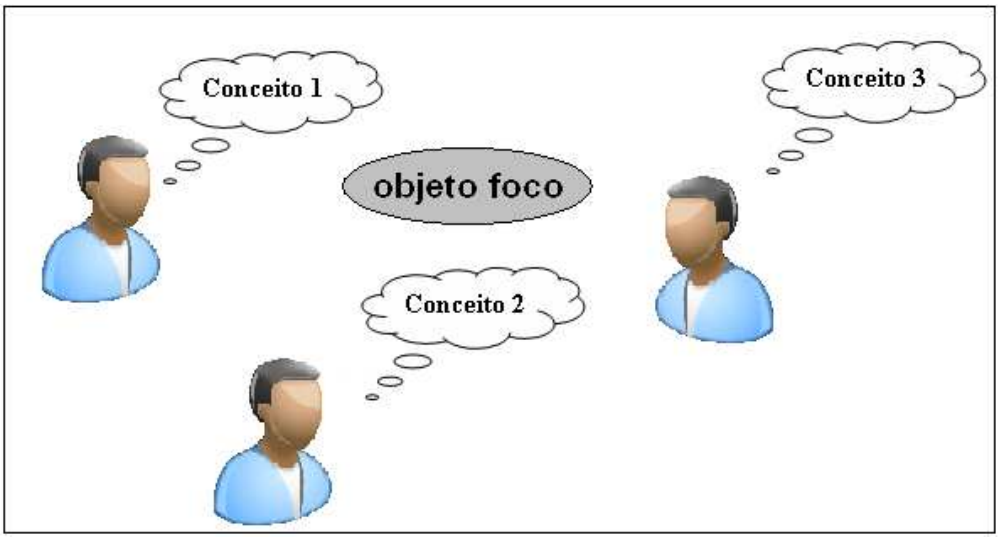


Figura 3.2: Compreensão do objeto foco

2. Captura do campo semântico ao redor deste objeto foco.

Figura 3.3: Capturando campo semântico

3. Transformação do campo semântico em uma *tag cloud*.

Figura 3.4: Processo de construção da Tag Cloud

4. Interpretação da *tag cloud* pelo usuário final.

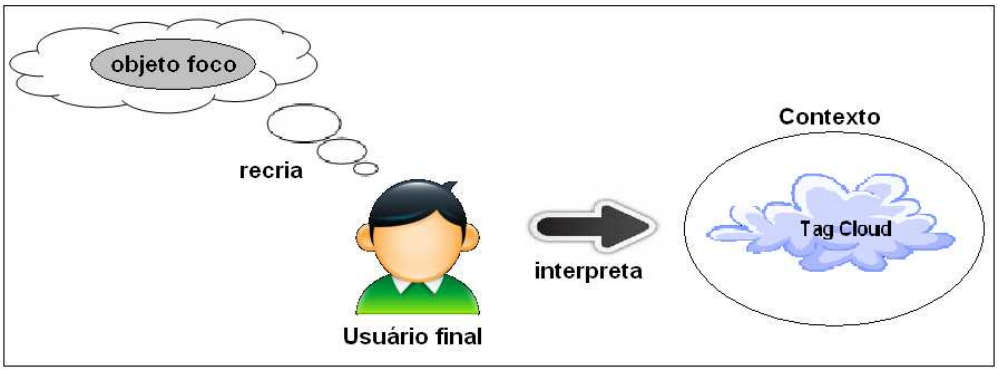


Figura 3.5: Interpretação da Tag Cloud

A quarta etapa deste processo de criação é o uso atual da *tag cloud*, a qual é "recriada" através da interpretação do usuário final que tenta entendê-la como um possível objeto dentro de um contexto. Ou seja, o usuário final interpreta a *tag cloud* e tenta recriar em sua mente o objeto que deu origem a esta *tag cloud*.

3.2 Propriedades Visuais

Quando se define qual será a apresentação adotada para uma *tag cloud* existem diversas propriedades que podemos considerar.

Para começar, em (Bateman *et al.*, 2008) é definida "influência visual" como: "as características visuais da *tag* que atraem a atenção dos usuários". Ainda é ressaltado que a influência visual da *tag* não inclui o valor semântico da mesma. Portanto, a principal tarefa ao se definir a apresentação de uma *tag cloud* é saber quais propriedades deverão ser utilizadas para dirigir a atenção dos usuários para as *tags* mais importantes.

Em (Bateman *et al.*, 2008) é apresentado um estudo realizado para descobrir quais propriedades são consideradas mais "visualmente importantes" pelos usuários. E por isso, os *designers* de *tag clouds* devem ter uma maior atenção com elas. Com o estudo realizado foi apresentado o seguinte resultado:

1. ***Propriedades visuais importantes***

Tamanho da fonte: tamanho da fonte utilizada na *tag*. Tem uma forte influência sobre os usuários e eles inclusive são capazes de reconhecer pequenas variações no tamanho.

Peso da fonte: peso utilizado na *tag* (negrito). Foi identificado como um importante fator visual. Foi indicada a utilização para dados do tipo binário e para destacar as *tags* principais.

Intensidade: intensidade das cores utilizadas na fonte da *tag*. Foi considerada uma boa propriedade para se obter a atenção visual. Foi recomendado alterar a intensidade da cor em intervalos de aproximadamente 10%, porém, deve se ter o cuidado para que em intensidades muito baixas não seja provocado nenhum problema para os usuários.

2. ***Propriedades visuais menos importantes***

Número de *pixels*: quantidade de *pixels* usados para formar a *tag*. As ações dos usuários não foram influenciadas por esta propriedade, podendo então ser ignorada pelos designers de *tag clouds*.

Largura da tag: As ações dos usuários também não foram influenciadas por esta propriedade, podendo então ser ignorada pelos designers de *tag clouds*.

3. ***Propriedades visuais para usar com atenção***

Cor: cor utilizada na fonte da *tag*. Foi percebido que as cores são facilmente identificadas pelos usuários, porém, não foi descoberto quais cores deveriam atrair mais facilmente a atenção dos usuários. Portanto, a idéia de como as cores serão utilizadas na *tag cloud* deve estar bem clara para eles, para que seja evitado uma interpretação equivocada.

Posição: forma como as *tags* serão distribuídas na *tag cloud*. Foi percebido que a parte central das *tag clouds* tende a ser a que obtém a maior atenção dos usuários e as partes superiores e inferiores as que obtêm a menor atenção. Por isso, foi recomendado que as *tags* que devem chamar mais atenção fiquem na parte central das *tag clouds*. Porém, nem sempre isso será possível, por exemplo, quando é utilizada uma *tag cloud* ordenada alfabeticamente.

Portanto, com o resultado apresentado por (Bateman *et al.*, 2008), propriedades como, o tamanho da fonte, peso da fonte e a intensidade devem ser utilizadas para representar a importância da *tag*. A cor da fonte e a posição das *tags* também são importantes, porém, estas propriedades devem ser utilizadas com cautela.

Em (Hassan-Montero, 2006) é apresentada uma idéia de posicionamento para as *tags* em que *tags* que possuem a mesma classificação semântica podem ser colocadas próximas umas das outras.

Em (Bielenberg, 2006) é apresentado um modelo de *tag cloud* em uma forma circular, na qual, o tamanho da fonte e a distância para o centro da *tag cloud* representam a importância de uma *tag*, porém, a distância entre as *tags* não representam suas similaridades. Na figura 3.6 é apresentado um exemplo desta *tag cloud*.

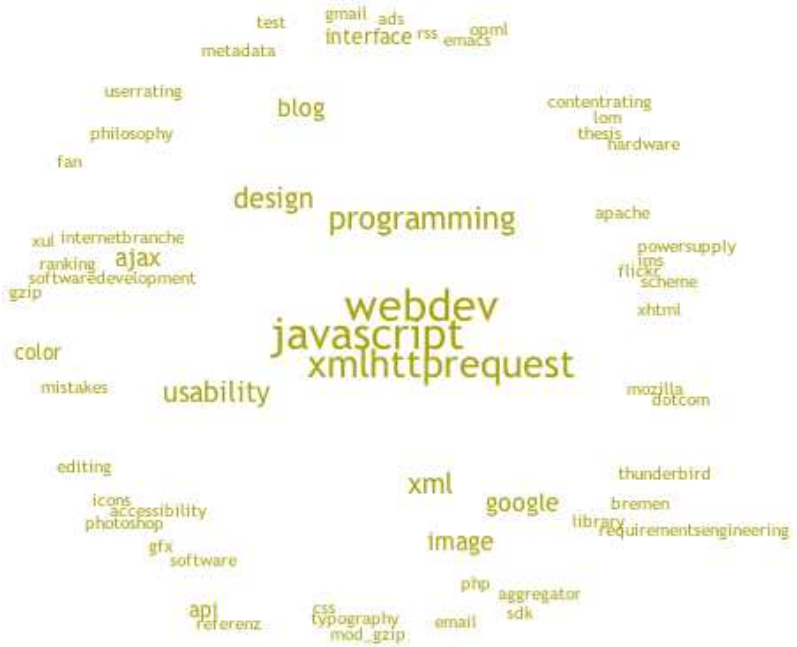


Figura 3.6: Exemplo do modelo de *Tag Cloud* proposto em (Bielenberg, 2006)

Apesar das primeiras gerações de *tag clouds* serem 2D, atualmente pode-se encontrar também *tag clouds* 3D, figura 3.7 [1] , o que provavelmente aumentará o número de propriedades a serem analisadas pelos *designers* de *tag clouds*.

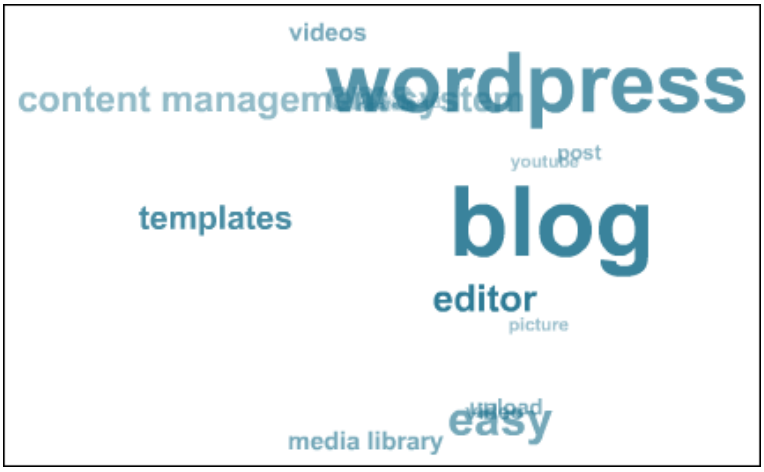


Figura 3.7: Exemplo de *Tag Cloud* 3D (http://www.cmswebsite.co.za/features/tag-cloud-3d/)

[1] http://www.cmswebsite.co.za/features/tag-cloud-3d/

3.3 Aplicações

Tag Clouds podem ser utilizadas em diversas tarefas, dentre as quais podemos citar tarefas que consistem na localização de um item ou conjunto de itens específicos e tarefas que tenham por objetivo fornecer uma visão geral sobre um determinado assunto. A seguir listamos algumas das tarefas as quais podem ser suportadas por *tag clouds* (Rivadeneira *et al.*, 2007):

- Busca: Localizar um termo específico ou algum que represente um conceito procurado, muitas vezes também permitindo a navegação para conteúdos adjacentes.
- Navegação: Utilização de *tag clouds* como meio de navegação, em certos casos sem um objetivo específico, sem item ou tópico em mente.
- Formar impressão: Utilizar *tag clouds* como meio para formar uma impressão ou visão geral do conjunto de dados ou da entidade a qual ela está associada. Esta impressão deve incluir tanto o conhecimento dos tópicos mais relevantes quanto o dos que aparecem em uma menor freqüência. Neste tipo de aplicação, tivemos o primeiro exemplo de sua utilização no resultado de um experimento conduzido pelo psicólogo social Stanley Milgram em 1976. Ele questionou algumas pessoas sobre lugares de Paris, e então criou um “mapa mental” coletivo da cidade usando o tamanho da fonte para indicar quantas vezes um determinado local havia sido mencionado, figura 3.8 (Viégas, 2008).
- Reconhecimento: Uso de *tag clouds* para o conhecimento e distinção de conteúdos.

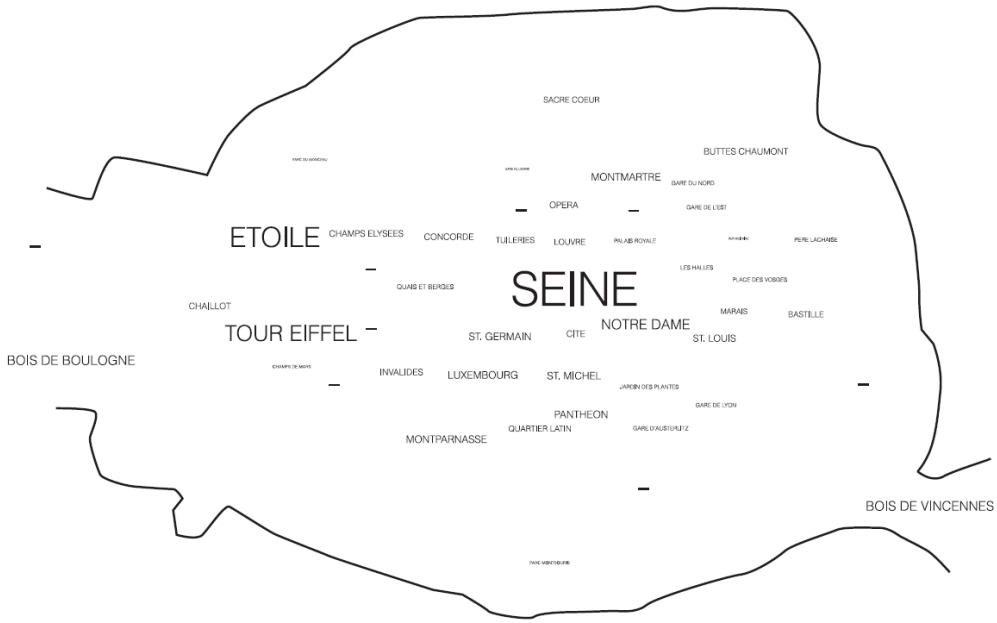


Figura 3.8: Mapa mental de Paris segundo Stanley Milgram, (Viégas, 2008)

- Identificação de similaridade entre conceitos: Através da análise de como usuários aplicam tags, como as tags são aplicadas aos links, e como os usuários coletam conteúdo, é possível calcular a "distância" entre tags, usuários e conteúdos. E isto pode se tornar uma poderosa ferramenta que permita que os usuários naveguem entre conteúdos (Shaw, 2005). E a forma de visualizar estas distâncias pode ser através de estruturas similares às *tag clouds*, figura 3.9.

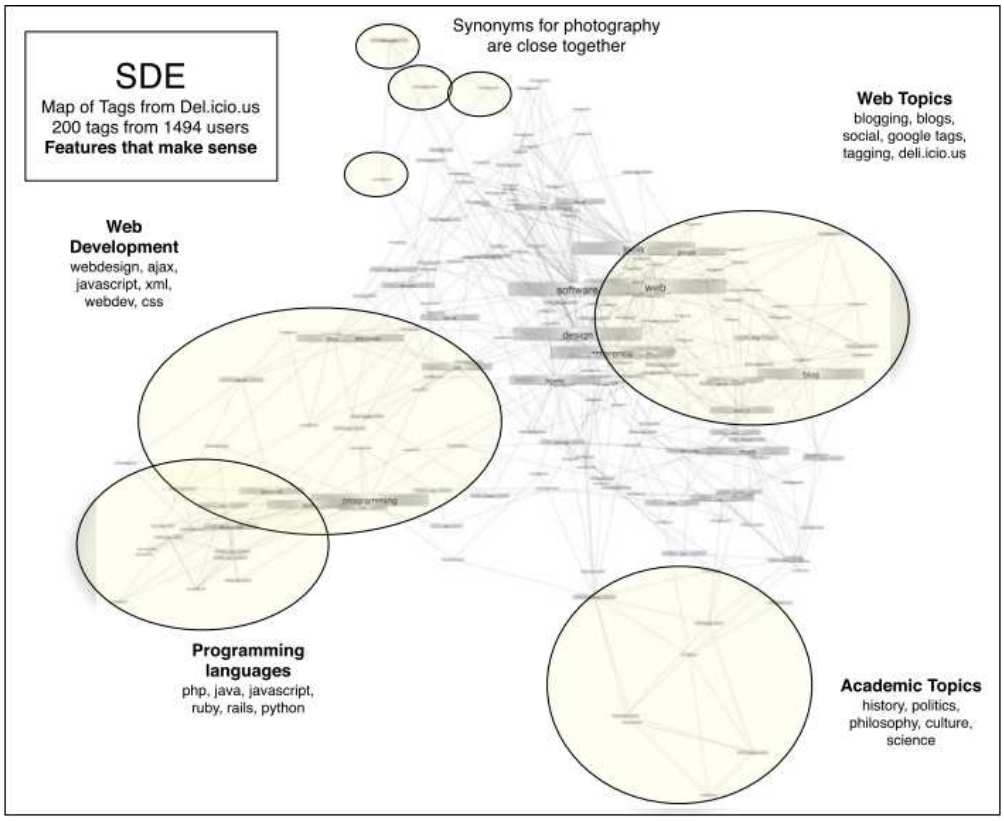


Figura 3.9: Representação da "distância" entre tags (Shaw, 2005)

- Visualizar evolução das *tags*: Em (Dubinko *et al.*, 2006) é apresentada uma abordagem para a visualização das *tags* mais utilizadas em um determinado intervalo de tempo. Nesta abordagem é permitido ao usuário observar e interagir com as *tags* que evoluem no decorrer do tempo.

3.3.1 Considerações sobre Tag Clouds

Nesta seção serão apresentados alguns resultados de estudos relacionados à utilização de *tag clouds*, assim como seus pontos positivos e negativos.

Hearst em (Hearst, 2008), apresenta em seu trabalho a possibilidade de que em certas situações a utilização de uma simples lista ordenada é melhor do que o uso de *tag clouds*. No entanto, ele ressalta que o uso de fontes de maior tamanho para destacar os termos mais importantes nas *tag clouds* tem um efeito positivo sobre as pessoas. Elas conseguiam relembrar com maior facilidade os termos com fontes maiores.

Neste mesmo trabalho após a realização de uma série de entrevistas são apresentadas algumas considerações sobre as *tag clouds* bem interessantes. As *tag clouds* foram consideradas úteis para mostrar tendências e também para exibir informação ou conteúdo dinâmico. Da mesma forma elas foram consideradas importantes para ilustrar qual é a essência do site, além de serem consideradas alegres e convidativas para que as pessoas possam interagir com o site que as usam.

Em (Viégas, 2008), temos a abordagem de uma aparente contradição que cercam as *tag clouds*, pois, se elas não provêem benefícios quantificáveis, e se as pessoas desconhecem como os itens são organizados, por que as *tag clouds* estão sendo tão utilizadas. Sites da *web* 2.0 tendem a atrair um grande número de usuários que contribuem com seus próprios conteúdos e tratar a diversidade e quantidade destas contribuições é um desafio. Portanto, ter *tag clouds* que resumam algumas destas atividades de uma maneira simples pode ser um recurso valioso para a comunidade de usuários. Ou seja, estas *tag clouds* podem agir como espelhos dos indivíduos e dos grupos. Além disso, esta visualização de todas as palavras é um diagrama que qualquer um pode gostar.

Seguindo na discussão, surge então a procura pelo motivo que faz as *tag clouds* serem utilizadas fora da esfera dos sites da *web* 2.0. Os autores então relataram que, através de uma experiência conduzida por eles no Many Eyes[2] eles perceberam que as *tag clouds* funcionam melhor como um retrato individual do que de grupos. E eles ainda ressaltam que, apesar dos problemas teóricos, *tag clouds* tornaram-se uma possível opção de ferramenta para a realização de análise.

---

[2] http://manyeyes.alphaworks.ibm.com/manyeyes/

Ainda de acordo com (Viégas, 2008), *tag cloud* é uma técnica que não veio da comunidade de visualização e que viola algumas das regras douradas de *design* visual tradicional. Apesar disso, a ampla popularidade das *tag clouds* e variedade de aplicações, desde experimentos psicológicos até análises de discursos de políticos, indicam que elas passaram no teste de aplicabilidade. Eles ainda ressaltam que alguém pode dizer que elas funcionam na prática, mas não na teoria.

Este aspecto falho segundo a visualização convencional merece atenção porque pode sugerir novas possibilidades. A demanda crescente por *tag clouds* indica que há uma importante classe de dados que os usuários querem visualizar, como, textos não estruturados. Além disso, a utilização desta forma de visualização como, sinalizador social ou ferramenta de análise textual, sugere que os especialistas em design de informação devem reavaliar os propósitos e objetivos de suas criações. Neste momento em que *designers* de fora da academia estão adotando técnicas de visualização acadêmicas, os teóricos podem retornar o favor e tomar inspiração a partir da explosão de criatividade fora da comunidade tradicional de *design* visual (Viégas, 2008).

Outro estudo conduzido por (Halvey, 2007) também destaca alguns pontos importantes. Entre estes pontos podemos citar:

- A organização por ordem alfabética pode ajudar os usuários a localizar o que eles procuram de forma mais rápida.
- O tamanho da fonte também é muito importante para que os usuários encontrem a informação desejada de forma mais rápida e fácil.
- O posicionamento das *tags* deve ser analisado com muita atenção.
- Destaca a característica de que os usuários têm mais facilidades para varrer superficialmente listas e *tag clouds* do que lê-las.

Também se estabelecermos uma comparação com padrões convencionais de navegação, como por exemplo, a utilização de *links*, listas ou menus, *tag clouds* não vão necessariamente oferecer uma navegação mais conveniente e intuitiva. No entanto, se for usada de uma forma adequada, ela pode mostrar rapidamente os tópicos, temas ou assuntos principais, fornecendo uma visão precisa sobre o conteúdo abordado pelo site. E como os seres humanos geralmente possuem a tendência de pensar em conceitos e modelos, é mais fácil apresentar o conteúdo desta forma, priorizando e destacando os conceitos mais importantes através de seus pesos (Friedman, 2007).

Portanto, podemos considerar que a principal vantagem no uso das *tag clouds* está na possibilidade de destacar os assuntos ou tópicos mais importantes e populares de uma forma dinâmica, algo que não é possível através dos padrões convencionais. Entre outras vantagens no uso das *tag clouds* podemos também destacar: seu design simples bem como seu fácil entendimento, além da necessidade de pouco espaço para que possa ser exibida. Em (Hearst, 2008) também é destacado que as *tag clouds* possuem a capacidade de mostrar a atividade mental e social das pessoas, e sua aparência reflete justamente isto.

Já entre as desvantagens relacionadas às *tag clouds*, pode-se destacar que palavras com um maior número de caracteres têm uma tendência natural a terem um maior destaque. Outra característica negativa das *tag cloud*s é a dificuldade de se comparar e analisar *tags* que possuam o mesmo tamanho.

Por isso, deve-se ter em mente que *tag clouds* não foram criadas para substituir os padrões convencionais de navegação, mas para serem utilizadas como uma alternativa ou ainda auxiliando a forma convencional de navegação.

3.4 Considerações Finais

Neste capítulo foram apresentadas algumas das definições estabelecidas sobre *tag clouds*. Além disto, foram identificadas suas propriedades principais, as aplicações que podem ser suportadas por elas e também foram citados seus prós e contras.

No capítulo a seguir será apresentado o modelo formal proposto para a construção de tag clouds.

4. Modelo formal para construção de *Tag Clouds*

Neste capítulo será apresentado o modelo proposto para a criação de *tag clouds*. Modelo que é um dos objetivos a ser alcançado por este trabalho e que permitirá continuar a conduzir a construção da ferramenta proposta. Este modelo será construído usando como base o modelo apresentado por (Marinchev, 2006).

4.1 Cenário atual

Atualmente nota-se o avanço de sites de natureza social, como Delicious[3], compartilhamento de *bookmarks*, Flickr[4], compartilhamento de fotos e Technorati[5], pesquisa de blogs. Também percebemos o aumento no incentivo ao uso de *tags* para descrever informações e características sobre os mais diversos objetos, desde documentos, fotos até músicas. Com isto, tornou-se comum apresentar estas *tags* em um formato conhecido como *"Tag Cloud"*.

De acordo com (Hassan-Montero, 2006), uma *Tag Cloud* é um conjunto de *tags* ordenados de tal forma, para que possam transmitir informação e significado através do uso de diferentes tamanhos de fonte, estilos e cores, baseados na sua importância dentro do grupo no qual elas aparecem. Quanto mais popular uma *tag* é, maior é a sua fonte e, portanto, mais destacada fica esta *tag* na *tag cloud* na qual ela aparece. Portanto, *tag clouds* fornecem um resumo ou uma visão (representação) semântica dos conceitos mais importantes para representar um objeto qualquer (Lamantia, 2006).

---

[3] http://delicious.com/

[4] http://www.flickr.com/

[5] http://technorati.com/

Seres humanos constroem este campo semântico associando os conceitos que são percebidos através da observação do objeto e as palavras que representam estes conceitos, isto por meio de sua interpretação particular. Se estes objetos são documentos, estas palavras podem ser encontradas em seus conteúdos ou inferidas através do entendimento do conteúdo do documento.

Geralmente, as pessoas podem facilmente identificar palavras para representar conceitos, desde que elas tenham um razoável conhecimento da linguagem utilizada e do mundo. Elas também podem descrever conceitos através de sentenças. No entanto, sistemas automáticos não possuem este conhecimento, portanto, eles devem inferir o conhecimento exclusivamente através das palavras que compõem o documento ou alguma outra informação textual, tais como metadados ou *tags* aplicadas a documentos similares.

Também devemos destacar que a relação entre palavras e conceitos não é bijetiva. Uma palavra simples pode representar mais do que um conceito, fato que é conhecido como polissemia. A palavra “manga”, por exemplo, pode representar uma fruta ou uma das partes que compõem uma camisa. Como outro exemplo, podemos citar a palavra “jaguar”, que pode se referir tanto ao animal quanto ao carro. Por outro lado, um conceito pode ser representado por várias palavras, tanto simples como compostas, tal como “carro”, “automóvel”, “veículo” e “meio de transporte”, o que é conhecido como sinonímia. Além disto, pode ocorrer de não existir palavra que solitariamente represente um conceito, como nos casos de, “banco de dados” e “banco de sangue”. Por último, também existem palavras que não carregam qualquer significado, tais como, preposições, artigos e conjunções, mas que ocorrem em grande freqüência em documentos textuais. Também é possível a existência de conceitos que não podem ser representado por palavras, mas isto é um caso muito raro. No entanto, com menor

raridade são encontrados conceitos que necessitam de textos complexos para que possamos compreendê-los.

Considerando estas limitações, uma solução automática para geração de *tag clouds*, semelhante a que é proposta neste trabalho, somente pode ser efetiva quando é desenvolvida como uma aproximação do comportamento humano. Para possibilitar esta aproximação, deve-se considerar que é necessário criar um modelo que permita a descrição do campo semântico de um documento, ambos quando analisados sob o ponto de vista humano e quando analisado sob o "ponto de vista" de um computador.

Nas seções a seguir, será iniciada a criação, desde o início, de uma definição conceitual e formal de *tag clouds* que irá permitir desenvolver um método abstrato para construí-las, método o qual será construído com base no modelo apresentado por (Marinchev, 2006).

4.2 Considerações iniciais

Neste trabalho é apresentado um modelo para geração de *tag clouds* a partir de documentos texto, tentando atingir ao máximo a aproximação do modelo de geração de *tag clouds* realizados pelos humanos. Para possibilitar isso foi decidido simular o processo descrito por (Marinchev, 2006), já apresentado anteriormente. No entanto, o primeiro passo do processo é desafiador, uma vez que, nele consta que para criar uma *tag cloud*, primeiramente, as pessoas devem criar conceitos em suas mentes. Estes conceitos são pensamentos abstratos que nem sempre podem ser descritos em palavras. Por exemplo, ao lermos o final da obra de Shakespeare, Romeu e Julieta, podemos notar um sentimento geral de tristeza, que só pode ser descrito aproximadamente por *tags* como "triste" e "infeliz".

Portanto, para adotar este processo deve-se primeiramente, decidir como representar conceitos em um computador. Vale destacar que esta representação não é a

mesma representação fornecida pelas *tags*. *Tags* são símbolos, normalmente palavras, que as pessoas podem entender e associar algum significado a elas. Já conceitos, no sentido cognitivo, são pensamentos abstratos, enquanto no sentido computacional eles devem ser modelados como alguma estrutura de dados, ou mesmo procedimento ou regra. Por exemplo, pode-se selecionar *Wordnet® synsets* (Fellbaum, 1998) para representar conceitos.

4.3 Definições iniciais

Esta seção começa apresentando alguns conceitos iniciais que serão úteis durante a definição do modelo.

Para começar é apresentada a definição de recursos e contextos. A motivação para estas definições está no fato de que, uma das propostas deste trabalho se baseia na construção de um modelo para geração de *tag clouds* que descrevam recursos em um determinado contexto.

4.3.1 Recurso e contexto

Recurso é qualquer conceito abstrato ou entidade física que pode ser identificado unicamente, seja na *web* ou fora dela (Berners-Lee *et al.*, 2005). A definição de recurso é deixada em aberto, para seguir a abordagem usada no RFC. Neste trabalho, um recurso é qualquer objeto identificável que pode ser descrito, ao menos parcialmente, por um conjunto de *tags*. Estas *tags* agem como representações dos conceitos que residem na mente das pessoas ou em estruturas de dados computacionais e podem ser aplicados aos recursos segundo alguma razão. Na *web*, recursos são identificados por *Uniform Resource Identifiers* (*URIs*). Existem muitas formas de representar propriedades de recursos, tal como metadados, no entanto a representação RDF (Manola, 2004) é padrão e estável. Um recurso é representado por uma letra *r*, possivelmente indexada.

Um **contexto**, denotado pela palavra *w*, é um conjunto de recursos que podem ser analisados como um todo. O contexto que contém todos os recursos será denotado por *W*, por causa de "web". Portanto, *w* não é um elemento de *W*, mas um subconjunto dele. Contextos podem ser abstratos, como quando são definidos por uma única palavra como "medicina", "engenharia", ou muito mais objetivo, tal como, quando é definido como "as respostas para a consulta 'jogo de futebol' feita a um específico sistema de busca". Contextos incluem documentos, objetos da vida real e eventos, ou seja, qualquer objeto que pode ser descrito como um recurso. Não existe obrigação para que todos os recursos de um contexto sejam do mesmo tipo.

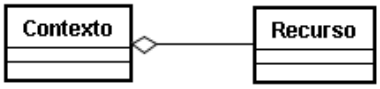


Figura 4.1: Modelo UML para recursos e contextos

4.3.2 Conjunto par atributo

Para começar, são definidos alguns conceitos que irão permitir estabelecer a definição do conjunto par atributo, que é uma abstração criada para definir um conjunto de atributos dinâmicos e seus valores para um objeto.

Um *objeto* é um conceito primitivo, portanto, não definido em teoria. Assim como na teoria orientada a objeto, objetos formam o conjunto raiz ao qual todos nossos outros conceitos definidos pertencem. Elementos atômicos e conjuntos de elementos pertencem ao conjunto de objetos. O conjunto de todos os objetos é o conjunto universal, denominado por *U*.

Um **domínio**, ou um **conjunto de valores**, representado por V_i, é um conjunto de valores. A idéia de domínio é usada tal qual na maior parte da teoria de banco de dados: definir um conjunto de possíveis valores para um atributo. Eles são indexados, como em V_i, para representar o fato que estamos usando múltiplos domínios. Valores em um domínio podem ser indicados por uma letra *v* duplamente indexada, tal como, $v_{ij,}$ para

mostrar que o valor v_{ij} pertence ao domínio V_i. Não temos qualquer pré-requisito sobre um domínio, por exemplo, ser finito ou composto de valores atômicos. O conjunto de todos os domínios é representado pela letra *V*, não indexada.

Um **atributo** *a* de um objeto *o* é uma propriedade que descreve *o*. O conjunto de todos os atributos possíveis é denotado por *A*. Apesar de várias vezes ser abstrato, atributos são normalmente representados nominalmente por *strings*. Portanto, espera-se que essa *string* seja uma palavra ou seqüência de palavras com algum significado claro.

Humanos em geral podem associar facialmente um domínio a um atributo, por exemplo, metros para avaliar distância e inteiros para medir idade. Por outro lado, programas de computadores precisam que esta associação seja feita explicitamente de alguma forma, como por exemplo, em um tipo de declaração semelhante às utilizadas em linguagens de programação.

Uma **função de atribuição de domínio** é uma função que associa um domínio com um atributo:

$$f_{ad}: A \rightarrow V.$$

Quando definido, uma função de atribuição de domínio representa os tipos de valores que podem ser atribuídos a um atributo.

Portanto, a partir deste momento consideramos que nossos conjuntos *A*, *V* e a função f_{ad} estão definidos. A seguir mostramos um exemplo com possíveis valores para os conjuntos e para a função:

$$A=\{cor,tamanho\}$$

$$V=\{Cores,SmallIntegers\}$$

$$Cores=\{vermelho,verde,azul,amarelo,preto,branco\}$$

$$SmallIntegers=\{1..256\}$$

$$f_{ad}=\{(cor,cores),(tamanho,SmallIntegers)\}$$

Um **par atributo** é um par ordenado (a_i,v_{ij}):

$$(a_i,v_{ij}) \in A \times V_i \text{ onde } f_{ad}(a_i)=V_i .$$

Pares atributos descrevem o valor de um atributo a_i, em um contexto particular. A definição de par atributo foi criada para permitir a seleção dinâmica de atributos que podem ser aplicados a um objeto. Desta forma, posteriormente, não será obrigatório definir previamente quais atributos podem ser usados para descrever um objeto, isto é, sua classe, como na teoria orientada a objetos.

Um **conjunto de par atributo**, ou um **mapa de tipo restrito** ou simplesmente um **mapa**, é um conjunto de pares atributos onde todo primeiro elemento de um par ordenado é único entre todos os componentes do mapa. Mapas serão utilizados futuramente para representar o conjunto de atributos que poderão ser utilizados para descrever um objeto. Um mapa será denotado por *m*, e definido formalmente como:

$$m = \{(a_i,v_{ij}) \mid (((a_i,v_{ij}) \in A \times V_i) \wedge (f_{ad}(a_i)=V_i)) \wedge ((\text{se } (a_i,v_{ij}) \in m) \wedge (\text{se } (a_k,v_{kn}) \in m)) \text{ então}$$

$$a_i=a_k \Rightarrow v_{ij}=v_{kn})\}$$

O conjunto de todos os mapas possíveis será denotado por *M*.

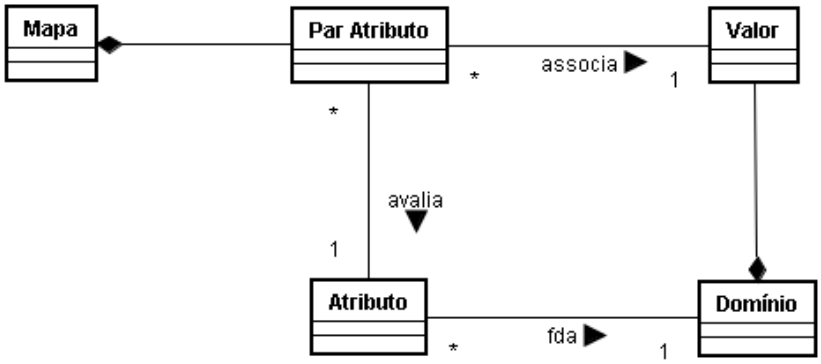


Figura 4.2 Modelo UML representando o modelo básico descrito

4.4 Funções de classificação e atribuição

Nesta seção serão utilizadas as definições do modelo básico para aplicar o conceito de atributos aos objetos.

Uma **função de classificação** é uma função f_c, que, dado um objeto *o* gera um conjunto de atributos A_i, $A_i \subset A$, que pode ser usado para representar os atributos do objeto.

$$f_c : O \rightarrow \wp(A)$$

$$f_c(o) = A_i = \{a_{ij} \mid a_{ij} \text{ é um atributo de } o\}$$

Uma **função de atribuição de mapa** é uma função f_{am} que, dado um objeto, e um conjunto de atributos A_i, $A_i \subset A$, gera um mapa no qual para cada atributo de A_i há um par atributo correspondente.

$$f_{am} : O \times \wp(A) \rightarrow M$$

$$f_{am}(o,A_i) = \{(a_{ij},v_{ij}) \mid \forall\, a_{ij} \in A_i \Rightarrow ((a_{ij},v_{ij}) \in f_{am}(o,A_i) \wedge (f_{ad}(a_{ij}) = V_i))\}$$

4.5 Aplicando atributos aos recursos

Nesta seção serão aplicados os conceitos de atributos aos recursos.

Uma **função de classificação de recurso** é uma função de classificação f_{cr} para a qual o conjunto de objetos é restringido ao conjunto de recursos.

$$f_{cr} : W \rightarrow \wp(A)$$

$$f_{cr}(r) = A_i = \{a_{ij} \mid a_{ij} \text{ é um atributo de } r\}$$

Uma **função de atribuição de mapa para um recurso** *r* é uma função de atribuição de mapa f_{amr} para a qual o conjunto de objetos é restringido ao conjunto de recursos. O mapa resultante geralmente representa as propriedades do recurso.

$$f_{amr} : W \times \wp(A) \rightarrow M$$

$$f_{amr}(r,A_i) = \{(a_{ij},v_{ij}) \mid \forall\, a_{ij} \in A_i \Rightarrow ((a_{ij},v_{ij}) \in f_{amr}(r,A_i) \wedge (f_{ad}(a_{ij}) = V_i))\}$$

Uma função de classificação de recurso é uma função que, dado um recurso, estabelece quais atributos podem ser avaliados para ele. Uma função de atribuição de mapa é uma função de avaliação que retorna os valores para um conjunto de propriedades de um recurso. Também pode ser entendido como a aplicação de um

conjunto de funções de avaliação onde cada uma retorna o valor de uma propriedade específica de um recurso.

A partir das definições expostas acima, temos agora o vocabulário para discutir como, dado um recurso, podemos dinamicamente gerar um conjunto de atributos e seus valores. Os conceitos descritos como conjuntos e funções podem ser vistos na figura 10 como um modelo UML.

Uma **representação de recurso** ou um **mapa de recurso** para um recurso *r*, *MR(r)*, é um mapa:

$$MR(r) = f_{amr}(r, f_{cr}(r))$$

Mapas de recurso agem como as representações dos recursos. Por exemplo, o mapa de recurso de um documento pode ser formado por tuplas que representem seu vetor de palavras.

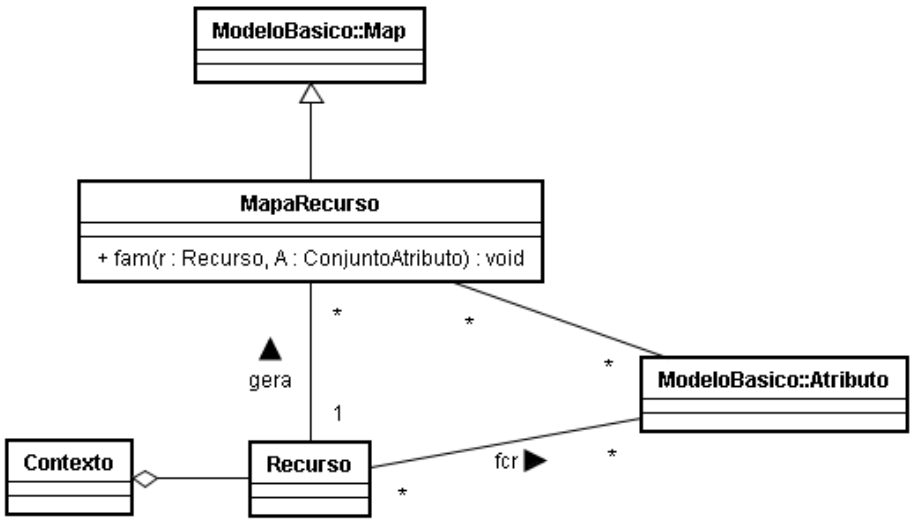


Figura 4.3: Modelo UML representando o uso de mapas para descrever recursos (como MapaRecurso)

4.6 Conceitos e Campos Semânticos

Nesta seção começamos a construir o conceito de uma *tag cloud* abstrata que poderia ser aplicada a qualquer objeto. Para começar, é formalizado o conceito de

campo semântico apresentado por (Marinchev, 2006). Para isto, deve-se supor a existência não só de um conjunto de recursos, mas também de um conjunto de conceitos, denotado por *C*. Conceitos podem ser extremamente abstratos como, por exemplo, no caso em que estamos falando sobre quando pessoas formam conceitos, ou bem mais concreto, como no caso da representação de conceitos em estrutura de dados.

Obter uma definição formal de conceito não é uma tarefa fácil, e tem gerado muitas discussões na filosofia. Neste trabalho é utilizada a abordagem de adotar a Teoria Clássica de Conceitos, segundo a qual: "a maioria dos conceitos são representações mentais estruturadas que codificam um conjunto de condições necessárias e suficientes para suas aplicações, se possível, em termos perceptivos e sensoriais" (Laurence, 1999). No entanto, é pretendido explicar conceitos tanto pelo aspecto humano quanto computacional. Portanto, vamos aceitar que conceitos não necessitem ser representações mentais, mas somente representações cognitivas adequadas.

Dado um recurso *r*, de um contexto *w*, e um conjunto de conceitos abstratos *C*, um **campo semântico** para *r* é um conjunto *CS(r)* de conceitos:

$$CS(r) = \{c_i \mid c_i \in C \wedge aplica(c_i,r)\}$$

onde, *aplica(c,r)* é um predicado lógico que representa o fato que um determinado conceito pode ser utilizado para descrever, de alguma forma, um recurso ou propriedade deste recurso.

Portanto, um campo semântico é um conjunto de conceitos abstratos que, de algum modo, podem ser aplicados a um recurso com o objetivo de construir algum entendimento sobre ele. Em alguns casos estaremos interessados em descrever o campo semântico de um recurso sob um específico contexto, e para representar isto será utilizado $CS_w(r)$.

Uma **função de classificação de conceito** é uma função de classificação f_{cc} que, dado um recurso *r* e um conceito *c* geram um conjunto de atributos A_i, $A_i \subset A$, que podem ser usados para representar os atributos do conceito *c* quando este se refere ao recurso *r*.

$$f_{cc} : W \times C \rightarrow \wp(A)$$

$$f_{cc}(r,c) = A_i = \{a_{ij} \mid a_{ij} \text{ é um atributo de c quando se refere a r}\}$$

Uma **função de atribuição de mapa para um conceito** *c* é uma função de atribuição de mapa f_{amc} que, dado um conceito *c*, um recurso *r* e um conjunto de atributos A_i, $A_i \subset A$, gera um mapa no qual cada atributo de A_i corresponde a um par atributo que o descreve.

$$f_{amc} : W \times C \times \wp(A) \rightarrow M$$

$$f_{amc}(r,c,A_i) = \{(a_{ij},v_{ij}) \mid \forall\, a_{ij} \in A_i \Rightarrow ((a_{ij},v_{ij}) \in f_{amc}(r,c,A_i) \wedge (f_{ad}(a_{ij}) = V_i))\}$$

Um **campo semântico avaliado** para um recurso *r* é um conjunto de pares ordenados no qual, o primeiro elemento é um conceito c_i aplicável ao recurso *r* e o segundo elemento é um conjunto par atributo composto dos atributos induzidos por c_i em *r*.

$$CSA(r) = \{(c_i,m_i) \mid c_i \in CS(r) \wedge m_i = f_{amc}(r,c_i,f_{cc}(r,c_i))\}$$

Apesar de (Marinchev, 2006) discutir somente campos semânticos, isto é, o mapeamento de conceitos aos recursos, neste trabalho é considerado que este mapeamento não pode ser livre de valores e informações adicionais. Além disso, computadores não lidam propriamente com conceitos, mas na verdade com alguma forma de representação de conceito que pode ser mapeado a ele. Estas representações possuem uma grande vantagem ao serem capaz de levar informação adicional consigo. Por exemplo, dado que escolhemos representar conceitos através de *synsets* do *Wordnet*, um *CS* para um recurso *r* pode ser o conjunto:

$$S = \{person,individual,someone,somebody,mortal,soul\}$$

No entanto, é interessante saber quais palavras foram usadas para obter o *synset*. Para isto, podemos ter um rótulo *palavras-originais* definindo um par atributo no nosso mapa para o *synset S*, e o conjunto *{person,individual}* descrevendo quais palavras foram encontradas no recurso *r* que gerou o *synset*.

Um **gerador de campo semântico** é uma função f_{gcs} que dado certo contexto *w*, um determinado recurso *r*, $r \subset w$, e um conjunto de conceitos, gera um campo semântico *CS(r)* o qual indica um conjunto de conceitos que podem ser considerados, sob alguma razão, serem aplicados a *r* em um contexto *w*.

$$f_{gcs} : W \times \wp(W) \rightarrow C$$

$$f_{gcs}(r,w) = \{c_i \mid c_i \in C \wedge aplica_w(c_i,r)\} = CS_w(r)$$

Um **gerador de campo semântico avaliado** é uma função f_{gcsa} que dado um contexto *w*, um específico recurso *r*, $r \subset w$, e um conjunto de conceitos, gera um campo semântico avaliado *CSA(r)* que indica um conjunto de conceitos que podem ser considerados, sob alguma razão, serem aplicados a *r* em um contexto *w* e seus mapas correspondentes.

$$f_{gcsa} : W \times \wp(W) \rightarrow C \times M$$

$$f_{gcsa}(r,w) = \{(c_i,m_i) \mid c_i \in CS_w(r) \wedge m_i = f_{amc}(r,c_i,f_{cc}(r,c_i))\}$$

Vale destacar que apesar de ser possível gerar um campo semântico a partir de um único recurso, é muito mais compreensível considerar que para gerar este campo semântico é necessário analisar não apenas o recurso, mas também o contexto no qual ele está inserido. Este contexto é caracterizado por todos os recursos que podem ser vistos de alguma forma como pertencentes ao mesmo escopo que o do recurso que está sendo analisado.

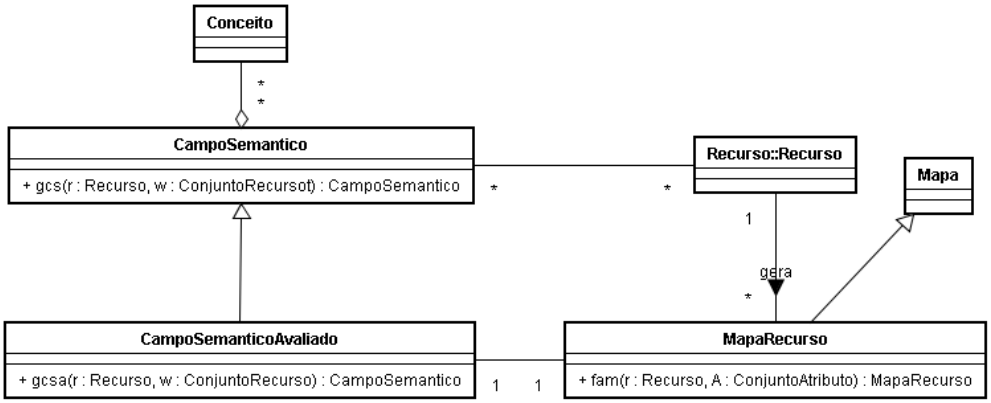


Figura 4.4: Modelando Campo Semântico em UML

4.7 Geração de *Tag Clouds* Abstratas a partir das *Tags*

Uma **função de classificação de *tag*** é uma função f_{ct}, que, dado um recurso *r* e uma *tag t* gera um conjunto de atributos A_i, $A_i \subset A$, que pode ser utilizado para representar os atributos da *tag t* quando se referem ao recurso *r*.

$$f_{ct} : W \times C \rightarrow \wp(A)$$

$$f_{ct}(r,t) = A_i = \{a_{ij} \mid a_{ij} \text{ é um atributo de t quando se refere a r}\}$$

Uma **função de atribuição de mapa para uma *tag t***, é uma função f_{amt} que, dado uma *tag t*, um recurso *r*, e um conjunto de atributos A_i, $A_i \subset A$, gera um mapa no qual para cada atributo de A_i há um par atributo correspondente o descrevendo.

$$f_{amt} : W \times C \times \wp(A) \rightarrow M$$

$$f_{amt}(r,t,A_i) = \{(a_{ij},v_{ij}) \mid \forall\, a_{ij} \in A_i \Rightarrow ((a_{ij},v_{ij}) \in f_{amt}(r,t,A_i) \wedge (f_{ad}(a_{ij}) = V_i))\}$$

Uma **representação de tag** ou um **mapa de tag** para uma tag t, *MT(t)*, é um mapa:

$$MT(t) = f_{amt}(r,t,f_{ct}(r,t))$$

Mapas de tag agem como as representações das *tags*. Por exemplo, o mapa de uma *tag* pode ser formado por tuplas que representam suas propriedades na *tag cloud.*

Dado um recurso *r*, de um contexto *w*, e um conjunto de *tags T*, um **conjunto de *tags*** para *r* é um grupo de *tags*

$$CT_w(r) = \{t_j \mid t_j \in T \wedge c \subset CS(r) \wedge representa(t_j,c)\}$$

onde cada *tag* t_j é um símbolo, normalmente uma palavra ou uma pequena seqüência de palavras, que representa um ou mais conceitos que podem ser aplicados a *r*.

Além disso, *representa(t,c)* é um predicado lógico que demonstra o fato que uma determinada *tag* pode ser utilizada para representar, de alguma forma, um conceito ou propriedade deste conceito.

Ressalta-se que não existe nenhum requisito sobre *CS(r)*.

Um conjunto de *tags* assume o papel de uma representação concreta de um campo semântico. Também, não existe diferença de um campo de *tags* criado por humanos ou computadores. Ambos são conjuntos de símbolos concretos. Um campo semântico pode induzir diferentes conjuntos de *tags* de acordo com os símbolos (palavras) disponíveis e a função *representa* escolhida.

Um **gerador de conjunto de *tags*** é uma função f_{gct} que dado um contexto *w*, um recurso específico *r*, $r \subset w$, gera um conjunto de *tags CT(r)* o qual representa o grupo de *tags* que podem ser consideradas, sob alguma razão, serem aplicadas a *r* no contexto *w*.

$$f_{gct} : W \times \wp(W) \rightarrow T$$

$$f_{gct}(r,w) = \{t_i \mid \exists\ c_i \in CS_w(r) \wedge representa_w(t_j,c_j)\} = CT_w(r)$$

Dado um conjunto de *tags CT(r)*, uma ***tag cloud* abstrata** *TCA(r)* é um conjunto de tuplas

$$TCA(r) = \{(t_j,m_i)\},$$

onde t_i é uma *tag* que pertence a *CT(r)*, e m_i é um mapa que representa os atributos desta *tag*.

Um **gerador de *tag cloud* abstrata** é uma função f_{gtca} que dado um contexto w, um recurso específico r, $r \subset w$, gera uma *tag cloud* abstrata $TCA(r,w)$ que indica um grupo de *tags* que podem ser consideradas, sob alguma razão, para representar os conceitos aplicáveis a r e seus mapas correspondentes.

$$f_{gtca} : W \times \wp(W) \rightarrow C \times M$$

$$f_{gtca}(r,w) = \{(t_i,m_i) \mid t_i \in CT_w(r) \wedge m_i = f_{amt}(r,t_i,f_{ct}(t,c_i))\}$$

Por exemplo, dado que o texto *"Romeu e Julieta"* é um recurso o conceito de *"amor proibido"*, o qual não pode ser representado a não ser por palavras, pode ser representado como duas *tags "proibido"* e *"amor"*, as quais podem aparecer na *tag cloud* abstrata como:

{(proibido,{(cor,preta),(tamanho,12),(x,1),(y,10)}),

(amor,{(cor,vermelha),(tamanho,16),(x,20),(y,20)})}.

Vale a pena ressaltar que nem todos os atributos devem ser representativos de características visuais. Possivelmente podemos ter atributos ocultos, tal como o valor do tf-idf (Manning *et al.*, 2008) da palavra no texto, o que poderá ser usado de alguma forma em uma representação ou algoritmo.

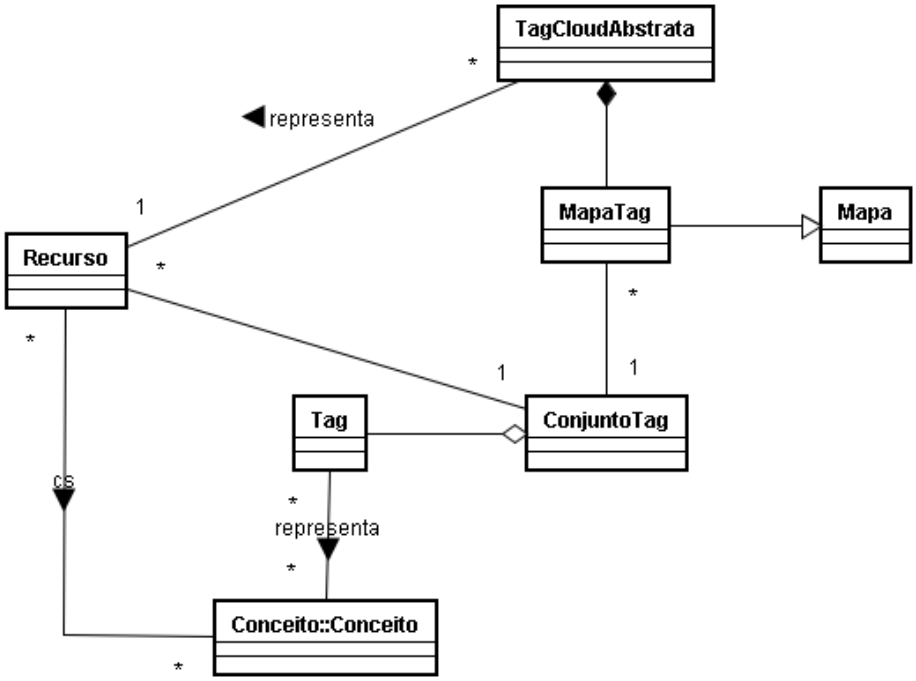


Figura 4.5: Modelando Tags e Tag Clouds Abstratas em UML

4.8 Criando Tag Clouds Abstratas

A partir das definições expostas, podemos desenvolver o processo de geração de *tag clouds* para um documento como:

1. Selecionar um documento e seu contexto;
2. Construir um campo semântico avaliado (CSA) para o documento;
3. Utilizar o campo semântico avaliado para definir uma *tag cloud* abstrata (TCA) para o documento;
4. Gerar uma representação visual para a *tag cloud* abstrata;

Não é levantada qualquer hipótese como este procedimento será desenvolvido. Estes detalhes são deixados em aberto e devem ser considerados particularmente em cada implementação. Por exemplo, não é estabelecido nenhum padrão de como fazer a representação de conceitos na *tag cloud*.

4.9 Tag Clouds propostas

Nesta seção são apresentados os modelos para as duas *tag clouds* que são propostas neste trabalho: ***Tag Cloud* de Resumo do Conjunto** e ***Tag Cloud* Diferencial**.

4.9.1 Modelo para uma Tag Cloud de Resumo do Conjunto

Nesta seção são exibidas as definições formais para o conceito de ***Tag Cloud* de Resumo do Conjunto**, *TCRC(w)*, que é apresentado neste trabalho. Seguindo o esquema das seções anteriores, é pretendido fornecer um modelo que possa ser usado por um processo para gerá-las.

Primeiramente, deve-se definir uma *TCRC(w)*, como uma *tag cloud* que resume os conceitos apresentados no conteúdo dos recursos que formam o contexto *w*.

Para começar, é apresentada a definição de ***Tag Cloud* de União do Conjunto**, *TCUC(w)*, a qual é obtida a partir da união de todas as *tag clouds* abstratas, de todos os recursos *r* que pertencem ao contexto *w*.

$$TCUC(w) = \bigcup_{r \in w} TCA(r)$$

A *tag cloud* de união do conjunto representa todos os recursos existentes em um determinado contexto. No entanto, não é necessariamente uma boa representação, uma vez que ela representa todos os conceitos possíveis de todos os recursos, mas não o resumo da coleção de recursos como um todo. E o nosso interesse é prover uma *tag cloud* que possa resumir a coleção de recursos de um contexto qualquer. Portanto, é necessária uma função que vá permitir determinar se um elemento desta *tag cloud* de união do conjunto pertencerá ou não a *tag cloud* que desejamos criar.

Uma **função de sumarização** é uma função $f_s(t)$, que dado uma tag *t* sabe decidir se ela deve ou não aparecer na *tag cloud* de resumo do conjunto.

$$f_s(t) \rightarrow \{true, false\}$$

Sendo assim, a definição da *TCRC(w)* é estabelecida como:

$$TCRC(w) = \{t \mid (t \in TCUC(w)) \wedge (f_s(t) = true)\}$$

Portanto, resumidamente esta *tag cloud* representa o conjunto de *tags* que estão relacionadas à pelo menos a algum dos recursos do contexto e, ao mesmo tempo, atendem às condições especificadas por $f_s(t)$. Geralmente, esta função utilizará algum *threshold* como filtro para selecionar somente as *tags* que irão pertencer a *tag cloud* do resumo do conjunto.

Desta forma, em uma *tag cloud* do resumo do conjunto irão constar as mais importantes *tags* para representar os conceitos existentes em uma coleção de recursos de acordo com algum critério.

4.9.2 Modelo para uma Tag Cloud Diferencial

Agora definimos formalmente o conceito de ***Tag Cloud* Diferencial**, *TCD(r,w)*. Seguindo o esquema das seções anteriores, pretende-se fornecer um modelo que possa ser utilizado por um processo para gerá-las.

De acordo com a proposta do trabalho também estamos interessados em como representar as diferenças de um recurso específico em relação aos outros integrantes da coleção. O conceito de diferença é usado no sentido de realçar, destacar ou evidenciar as características ou conceitos particulares de um recurso que não estejam sendo abordados pelos demais recursos neste mesmo contexto.

Portanto, uma *TCD(r,w)* é definida como uma *tag cloud*, que irá fornecer uma representação visual dos conceitos presentes em um específico recurso *r*, e que de alguma forma, se destacam ou se diferenciam dos demais conceitos presentes nos outros recursos, que são componentes de um determinado contexto *w*.

Para começar, é apresentada a **função de diferenciação** que é uma função $f_d(t,r,w)$, que irá fornecer uma condição para definir se a tag *t* deverá ser utilizada para representar um conceito que existe no recurso *r* que está sendo analisado mas não aparece na TCRC(w).

$$f_d(t,r,w) = \{true, false\}$$

Desta forma, também podemos definir uma *tag cloud* diferencial, *TCD(r,w)* como:

$$TCD(r,w) = \{ t \mid (t \in TCRC(w) \vee t \in TCA(r)) \wedge fd(t,r,w) = true \}$$

Logo, na *TCD(r,w)* são encontradas *tags* que descrevem os documentos individualmente, mas não são necessariamente consideradas na *tag cloud* de resumo do conjunto.

Sendo assim, a principal função de uma *tag cloud* diferencial é representar características individuais e importantes de cada recurso e por consequência auxiliar aos usuários no momento em que estes realizam uma busca, em certo contexto, a encontrar os recursos que eles desejam.

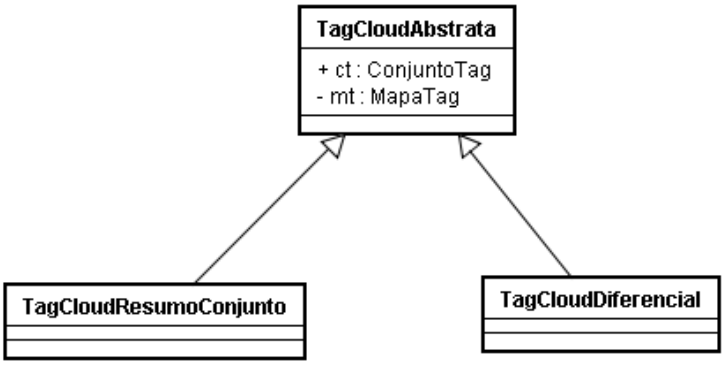


Figura 4.6: Generalização entre as Tag Clouds apresentadas

4.10 Considerações Finais

Neste capítulo foi apresentado o modelo proposto para a criação de *tag clouds*, assim como a definição para os dois novos tipos de tag clouds propostas. Este modelo

vai ser utilizado nas seções seguintes para que seja possível construir a ferramenta que também é proposta neste trabalho.

5. Métodos e Implementação

5.1 Base de teste utilizada

Para testar a geração das *tag clouds* propostas neste trabalho, optou-se por gerá-las em cima de uma base de teste preparada manualmente. Isto foi feito para que em cima de uma base mais limpa e criada especialmente para esta finalidade fosse possível obter melhores resultados e desta forma explicar de um melhor modo seus objetivos e funcionamento.

Com este objetivo foram criadas duas bases de texto uma em inglês e outra em português. A base em inglês foi criada em cima do termo "Jaguar" e abrangia os assuntos relacionados ao animal Jaguar, ao fabricante de carros de marca Jaguar, a jogos eletrônicos desta mesma marca entre outros. A base para o português foi criada sobre o assunto "Banco de Dados" e cobria temas e conceitos relacionados a esta área. Nas tabelas 5.1 e 5.2 são apresentados um resumo com estas informações.

	Animal	Carro	Video Game	Esporte	Total
Jaguar	17	19	5	2	43

Tabela 5.1: Distribuição de assuntos na base sobre Jaguar

	SGBD	Dados	Categorias	Modelagem	Outros	Total
Banco de Dados	7	3	3	4	6	23

Tabela 5.2: Distribuição de assuntos na base sobre Banco de Dados

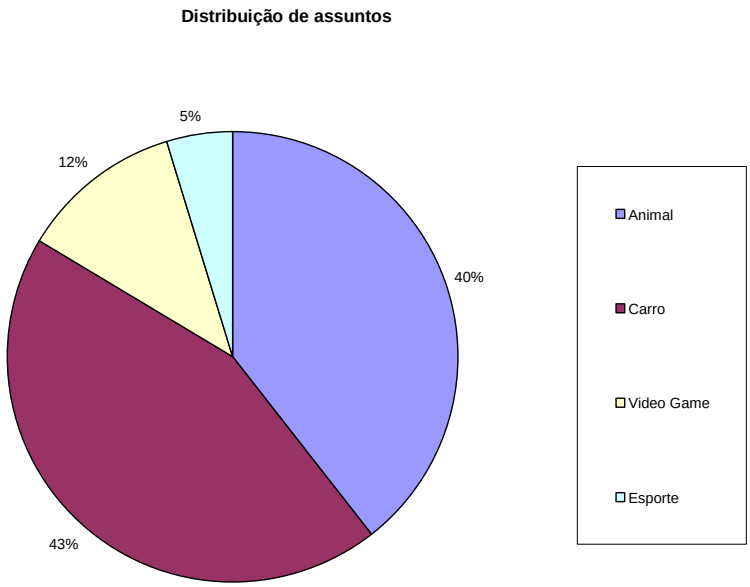


Figura 5.1: Distribuição de assuntos sobre o assunto Jaguar

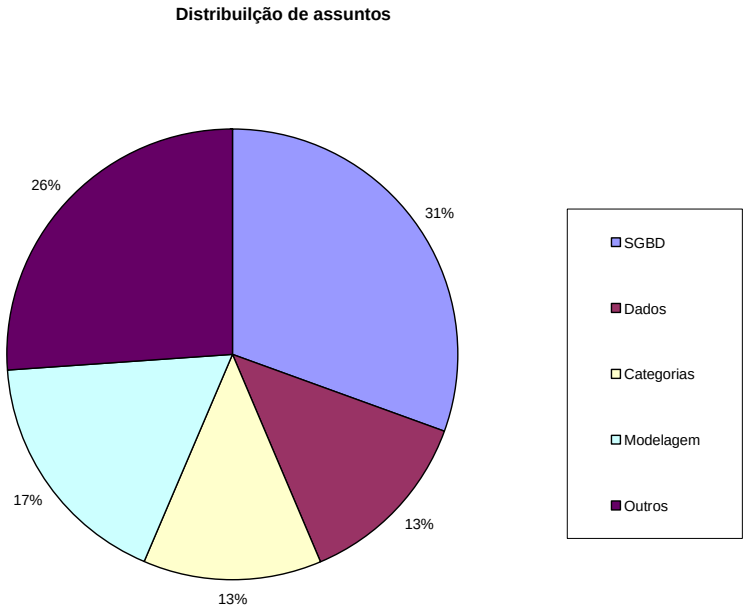


Figura 5.2: Distribuição de assuntos sobre o assunto Banco de Dados

5.2 Implementação

A ferramenta começa realizando todo o pré-processamento necessário para os documentos da base, como descrito em (Weiss *et al.*, 2005). Entre as etapas de pré-processamento que a ferramenta realiza estão: a *tokenização* do texto, eliminação das *stopwords*, *stemmização* e por último é aplicado um *POS-Tagger* para manter apenas os substantivos.

Estas etapas do pré-processamento foram implementadas de tal forma que fosse possível tratar tanto textos de língua inglesa quanto textos em português.

Em seguida são apresentadas as ferramentas utilizadas para realizar estas tarefas para cada um dos idiomas.

Primeiramente, serão apresentadas as ferramentas para tratar os documentos em inglês. Para realizar a *tokenização* do texto e criar os vetores de palavras para representar os documentos utilizamos a biblioteca Java, Word Vector Tool (WVTool)[6]. Na tabela 5.3 são apresentados alguns dos termos da lista de *stopwords* utilizada para a língua inglesa. Para o processo de *stemming* foi utilizado o conhecido algoritmo de Porter (Porter, 1980). E o *Part-Of-Speech Tagging* utilizado foi o desenvolvido pelo Grupo de Processamento de Linguagem Natural de Stanford[7].

a	but	what
after	can	when
again	did	which
almost	do	who
always	does	will
an	...	with
and	until	without
are	very	would
before	was	yet
below	were	your

Tabela 5.3: Exemplo de stopwords para a língua inglesa

[6] http://sourceforge.net/projects/wvtool/

[7] http://nlp.stanford.edu/software/tagger.shtml

Já para tratar os documentos em português foram utilizadas as seguintes ferramentas. Também foi utilizado o WVTool para realizar o pré-processamento do texto. Na tabela 5.4 são apresentadas algumas das *stopwords* utilizadas para o português. Para a tarefa de *stemming* foi utilizado o PTStemmer[8], uma biblioteca Java que possui uma implementação do algoritmo de *stemming* de Porter para a língua portuguesa. Para realizar o *Part-Of-Speech Tagging* para português foi utilizado o TreeTagger for Java (TT4J)[9], que é um *wrapper* para o popular *Part-Of-Speech tagger* TreeTagger[10], o qual pode ser utilizado para rotular textos em português.

à	cada	se
às	com	sob
a	da	sobre
agora	de	são
ainda	depois	também
algum	desde	tudo
ambos	dessa	um
ao	desta	vez
as	destes	vos
até	...	vão

Tabela 5.4: Exemplo de stopwords para a língua portuguesa

Para cada um dos idiomas também é realizado o enriquecimento dos vetores de palavras através da geração de *bigrams* (grupos de duas palavras).

Após a realização deste processamento, são obtidos os vetores de termos que representam cada documento da base. Em seguida, é calculado o valor do *Tf-idf* para cada um dos termos componentes dos vetores. *Tf-idf* é uma medida tradicional de relevância do termo utilizada em *information retrieval*, e definida como (Manning *et al.*, 2008):

$$w_{ij} = \frac{tf_{ij}}{\max_k(tf_{ik})} \times \log(\frac{N}{n_j})$$

onde

w_{ij}: é o peso do termo j no documento i

[8] http://code.google.com/p/ptstemmer/

[9] http://www.annolab.org/tt4j/

[10] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

tf_{ij}: é a frequência do termo j no documento i

N: é o número de documentos na coleção

n_j: é o número de documentos com o termo j

Outras medidas também foram avaliadas para a utilização, mas ficaram fora deste trabalho, como por exemplo, *information gain* (Dasgupta et al., 2007). Neste trabalho é utilizado "termo" quando é pretendido se referir ao radical de uma palavra qualquer. Também é mantida uma lista que relaciona todos os termos a suas palavras de origem.

Nas seções a seguir serão abordados os métodos que foram utilizados para gerar as *tag clouds* propostas neste trabalho.

5.2.1 Primeira abordagem

Nesta primeira abordagem as *Tag Clouds* Abstratas (TCA) de cada documento foram geradas seguindo o método apresentado a seguir.

Após a obtenção dos valores de *Tf-idf* de todos os termos de um determinado documento, organizaram-se estes termos em um ranking de acordo com estes valores. Deste ranking selecionamos no máximo 40 termos para compor o nosso Campo Semântico Avaliado (CSA). É importante ressaltar que ao construir o CSA eram recuperadas as palavras que originaram cada termo.

E então, a partir do CSA era gerada a TCA do respectivo documento. Deste conjunto de palavras são retiradas as *tags* que irão compor a TCA. Entre as propriedades que são utilizadas para gerar a TCA estão: tamanho, cor, posição, etc. Finalmente, com a TCA construída cria-se uma representação visual para a mesma.

Em seguida, é criada a *Tag Cloud* de Resumo do Conjunto (TCRC). Para isto primeiramente é gerada a matriz de termo x documento para a base. Uma matriz termo x documento é uma matriz na qual, cada linha representa um termo que existe na coleção e cada coluna equivale a um documento. Nela cada célula a_{ij}, representará a quantidade de vezes que o termo i ocorreu no documento j.

Na seqüência, em cima desta matriz são aplicadas as técnicas de Análise Semântica Latente (LSA), para descobrir a correlação entre os termos existentes na base. Para fazer a implementação do LSA foram utilizadas algumas das funcionalidades oferecidas pelo pacote Java Matriz Package (JAMA)[11]. Entre elas estão:

- Operações sobre matriz multiplicação, transposição, etc.
- Decomposição em valores singulares (SVD).

Após a obtenção da matriz reconstruída ao final do LSA calculamos a correlação de Pearson (Rodgers, 1988) entre os termos representados nesta matriz. Com as correlações calculadas, foi decidido construir o CSA da seguinte forma: utilizou-se como uma espécie de raiz, o termo utilizado na consulta. A partir dele obtemos os termos com maior grau de correlação que estão relacionados a ele e que possuam este grau maior que 0. Então estes termos são recuperados até atingir um limite de no máximo 40 termos. Além disso, também são selecionados alguns *bigrams* com um alto grau de correlação, mesmo que fora deste limite de 40, e que possuam entre os termos que formam o *bigram* algum dos termos utilizados na consulta. Como anteriormente, ao construir o CSA são recuperadas as palavras que deram origem a cada um dos termos.

Na seqüência a partir do CSA é gerada a TCRC. Seguindo o modelo anterior, após definido o conjunto de palavras temos as *tags* que irão compor a TCRC. Para criar a representação visual da TCRC são utilizadas as mesmas propriedades utilizadas ao gerar a TCA como: tamanho, cor, posição, etc.

Já para construir a *Tag Cloud* Diferencial (TCD) de cada documento, nesta primeira abordagem foi decidido utilizar as mesmas *tags* que compõem a TCA do respectivo documento. Isto porque, segundo o objetivo da TCD definido nas seções anteriores, o qual é destacar as características referentes ao documento que o diferenciam dos demais documentos do conjunto, entendeu-se que as *tags* formadoras da TCA estavam cumprindo este requisito

[11] http://math.nist.gov/javanumerics/jama/

satisfatoriamente. Para criar a representação visual da TCD são utilizadas as mesmas propriedades das demais *tag clouds* criadas.

Algoritmo Primeira Abordagem

Entradas

Coleção de documentos D={D1,D2,D3,...}

Criar TCA

Início

1:Para cada documento d petencente a D
2:Calcular tf-idf dos termos
3:Organizar ranking R dos termos pelo tf-idf
4:Selecionar um subconjunto de R para formar o CSA
5:Recuperar palavras que deram origem aos termos do CSA
6:Gerar a TCA a partir do CSA

Fim

Criar TCRC

Início

1:Gerar matriz M de termos x documentos
2:Aplicar o LSA na matriz M
3:Calcular a correlação de Pearson entre os termos da matriz
4:Selecionar termos com maior correlação
5:Gerar CSA da TCRC
6:Recuperar palavras que deram origem aos termos do CSA
7:Gerar a TCRC a partir do CSA

Fim

Criar TCD

Início

1:Utilizar as TCA geradas

Fim

5.2.1.1 Análise de Resultados

Nesta seção serão apresentados e analisados alguns dos resultados que foram obtidos nesta primeira abordagem.

Primeiramente, foi utilizada com a ferramenta uma base na língua inglesa. Para isso, utilizou-se a base de teste sobre Jaguar. Em seguida é apresentada a TCRC gerada.

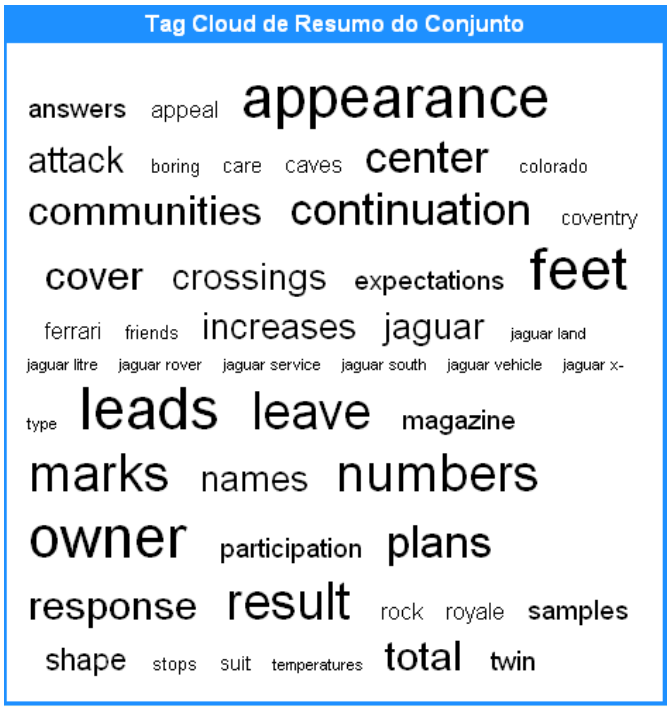


Figura 5.3: TCRC sobre a base Jaguar

Nesta TCRC é possível perceber tags que descrevem o conteúdo da base utilizada para esse primeiro teste, como “jaguar”, “communities”, “caves”, etc.

Em seguida são mostradas 3 TCD geradas para 3 diferentes documentos existentes nesta base.

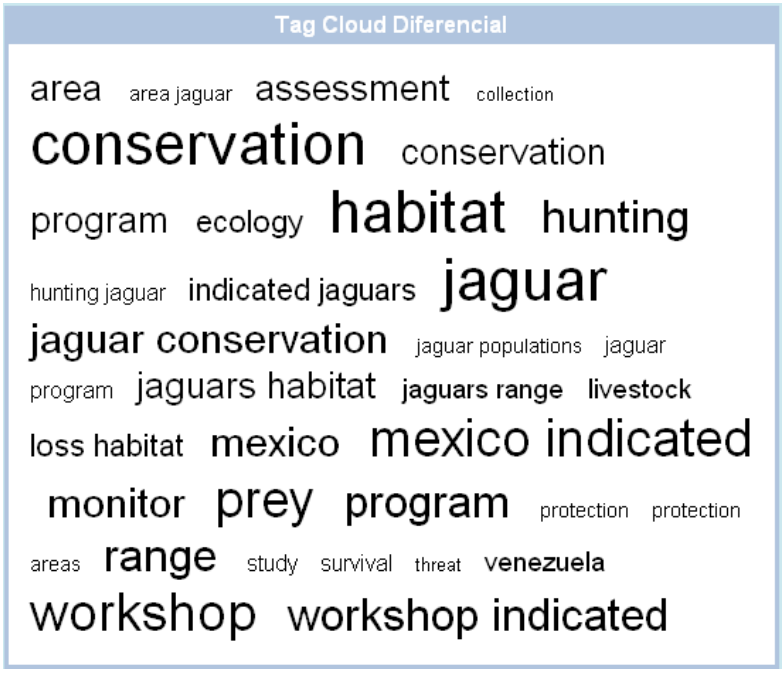


Figura 5.4: TCD de um documento referente ao animal Jaguar

Na figura 5.4 é apresentada uma TCD sobre um documento[12] que trata sobre um programa de conservação dos jaguares. É possível notar *tags* na TCD que descrevem razoavelmente bem o assunto tratado no documento, tais como, “conservation”, “habitat”, “jaguar conservation”, “program”, etc.

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

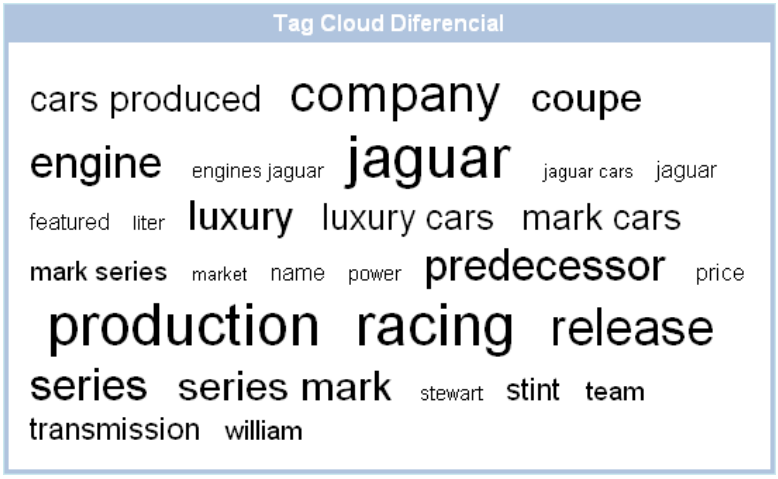


Figura 5.5: TCD de um documento referente ao fabricante de automóveis Jaguar

Na figura 5.5 é apresentada uma TCD sobre um documento[13] acerca do fabricante de carros Jaguar. Pode-se notar *tags* na TCD que estão relacionadas ao tema do documento, como por exemplo, “cars”, “engine”, “luxury cars”, “production”, “racing”, etc.

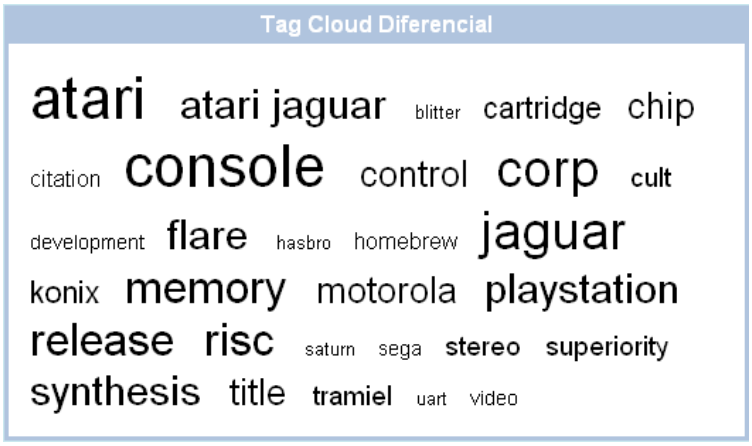


Figura 5.6: TCD de um documento referente ao console Jaguar

Na figura 5.6 é apresentada uma TCD sobre um documento[14] acerca do console de vídeo game jaguar fabricado pela Atari. Como nas outras TCD pode-se perceber *tags* na TCD

[13] http://www.jaguar-automotives.com/

[14] http://en.wikipedia.org/wiki/Atari_Jaguar

que estão relacionadas ao assunto deste documento, como, “atari”, “cartridge”, “console”, “video”, etc.

Na sequência, para o segundo teste foi utilizada com a ferramenta uma base na língua portuguesa. Para isso, foi utilizada a base de teste sobre Banco de Dados. Em seguida é apresentada a TCRC gerada.

Figura 5.7: TCRC sobre a base Banco de Dados

Nesta TCRC percebem-se *tags* que descrevem o conteúdo da base utilizada para esse primeiro teste, entre elas, “dados”, “modelagem”, “sgbd”, etc.

Em seguida mostramos 3 TCD para 3 documentos existentes nesta base.

Figura 5.8: TCD de um documento referente conceitos sobre Banco de Dados

A figura 5.8 apresenta uma TCD sobre um documento[15] que aponta alguns dos pontos importantes referente ao assunto de banco de dados. É possível perceber na TCD algumas *tags* que representam estes temas, como, “armazenamento”, “base dados”, “modelo”, “transação”, etc.

[15] http://pt.wikipedia.org/wiki/Banco_de_dados

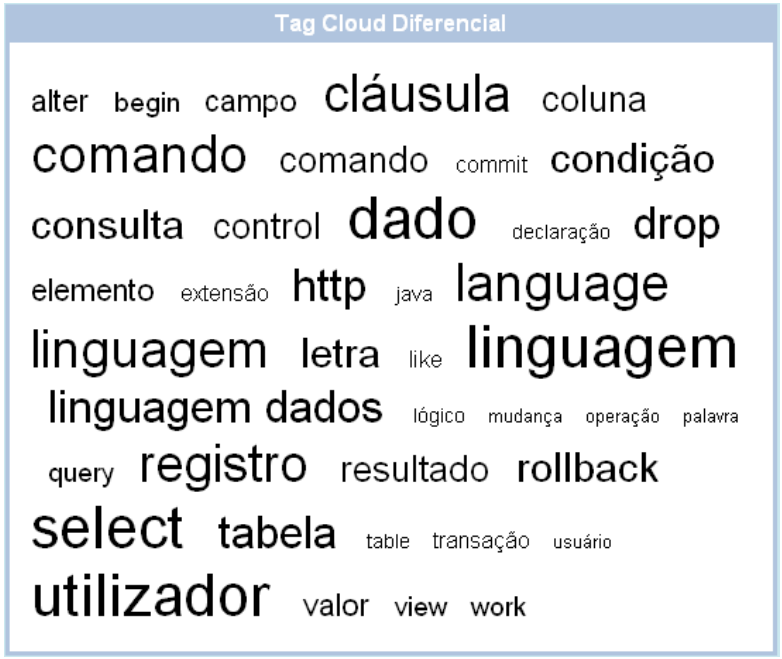


Figura 5.9: TCD de um documento referente à SQL

Na figura 5.9 é exibida uma TCD sobre um documento[16] que aborda a linguagem SQL. Pode-se notar, *tags* na TCD que estão relacionadas ao tema do documento, como, “cláusula”, “comando”, “linguagem”, “registro”, “resultado”, “select”, etc.

[16] http://pt.wikipedia.org/wiki/SQL

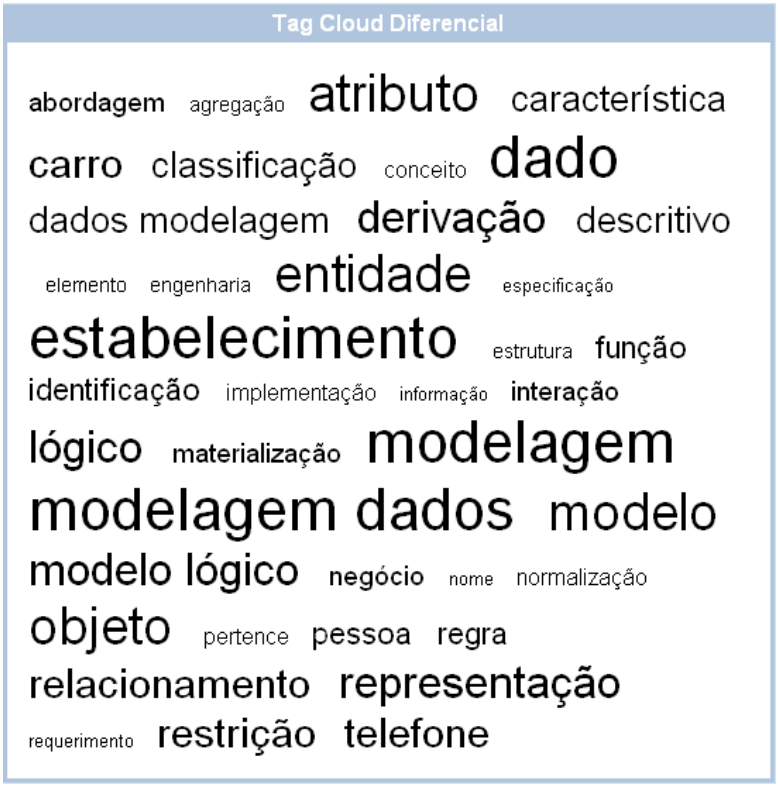


Figura 5.10: TCD de um documento referente à modelagem de dados

Na figura 5.10 é mostrada uma TCD sobre um documento[17] que trata sobre o tema de modelagem de dados. Pode-se notar *tags* na TCD que descrevem o assunto abordado pelo documento, como por exemplo, “atributo”, “dado”, “entidade”, “estrutura”, “modelagem dados”, “representação”, etc.

[17] http://pt.wikipedia.org/wiki/Modelo_de_dados

5.2.2 Segunda abordagem

Nesta segunda abordagem foram geradas as *Tag Clouds* Abstratas (TCA) de cada documento através de um método muito semelhante ao qual foi utilizado para gerá-las na primeira abordagem.

A principal diferença fica no seguinte ponto. Após a construção do CSA, seguindo o processo utilizado anteriormente, ele é enriquecido adicionando conceitos que estejam relacionados aos termos que estão presentes nele. Estes conceitos são selecionados a partir da seleção de *synsets* do Wordnet[18].

Como anteriormente, simultaneamente à criação do CSA são recuperadas as palavras que deram origem a cada um dos termos. Na seqüência, a partir do CSA é gerado a TCA do respectivo documento. Nesta segunda abordagem mantivemos as mesmas propriedades para criar a representação da TCA.

Para criar a *Tag Cloud* de Resumo do Conjunto (TCRC) foi utilizada a seguinte metodologia. Foi organizado um ranking com todos os termos presentes em cada um dos CSA que foram criados para cada documento do conjunto. Então os termos foram ordenados de acordo com a quantidade de CSA que cada um deles está presente. E em seguida, foram selecionados os 40 primeiros termos para compor o CSA da TCRC.

Assim como anteriormente, ao mesmo tempo em que é criado o CSA são recuperadas as palavras que deram origem a cada termo. Com a obtenção destas palavras, estão definidas as *tags* que irão compor a TCRC.

Para gerar a *Tag Cloud* Diferencial (TCD), são selecionadas dentre as *tags* da TCA do respectivo documento aquelas que não estão presente na TCRC, até um limite de 40 *tags*.

Para criar a representação visual da TCRC e da TCD são utilizadas as mesmas propriedades que foram utilizadas nas outras *tag clouds* criadas. É importante destacar que

[18] http://wordnet.princeton.edu/

nas TCD as *tags* originadas a partir de conceitos do *Wordnet* são apresentadas em uma coloração diferente das demais *tags*.

Algoritmo Segunda Abordagem

Entradas

Coleção de documentos D={D1,D2,D3,...}

Criar TCA
Início

1:Para cada documento d petencente a D
2:Calcular tf-idf dos termos
3:Organizar ranking R dos termos pelo tf-idf
4:Selecionar um subconjunto de R para formar o CSA
5:Recuperar palavras que deram origem aos termos do CSA
6:Inserir conceitos no CSA a partir de synsets do Wordnet
7:Gerar a TCA a partir do CSA

Fim

Criar TCRC
Início

1:Criar lista com todos termos presente nos CSA dos documentos de D
2:Ordenar termos pela quantidade de CSA em que aparece
3:Selecionar termos com maior ocorrência
4:Gerar CSA da TCRC
5:Recuperar palavras que deram origem aos termos do CSA
6:Gerar a TCRC a partir do CSA

Fim

Criar TCD
Início

1:Para cada documento di petencente D
2:Selecionar tags da TCA que não estão presentes na TCRC
3:Gerar TCD a partir destas tags

Fim

5.2.2.1 Análise de Resultados

A seguir são apresentados alguns dos resultados obtidos nesta segunda abordagem.

Inicialmente a ferramenta foi utilizada com uma base na língua inglesa. Para possibilitar uma padronização nos resultados apresentados neste trabalho, foi utilizada a mesma base de teste sobre Jaguar que foi utilizada anteriormente. Em seguida é apresentada a TCRC gerada para esta base.

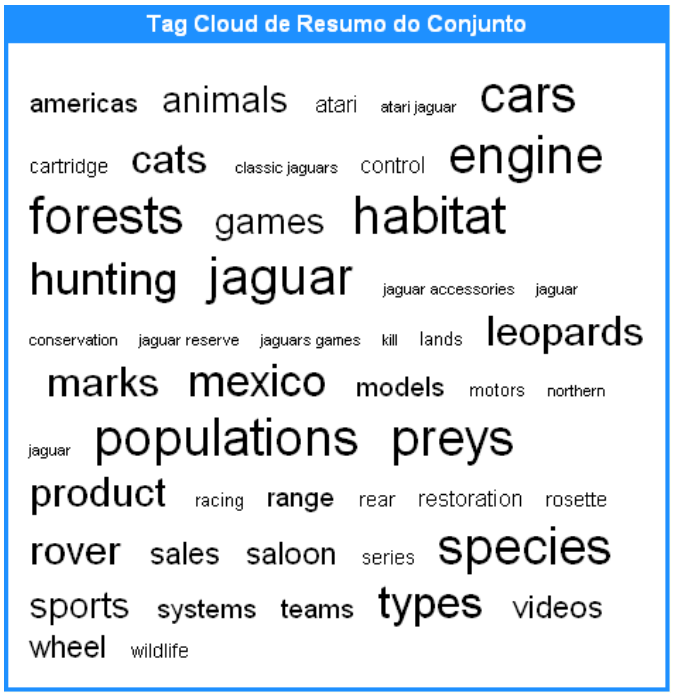


Figura 5.11: TCRC sobre a base Jaguar

Facilmente pode-se notar nesta TCRC *tags* que fornecem um resumo do conteúdo tratado na base utilizada para esse primeiro teste, como “animals”, “atari”, “cars”, “engine”, “games”, “sports”, etc.

Em seguida são exibidas as 3 TCD geradas para os mesmos 3 documentos, na mesma ordem, que também foram utilizados anteriormente como exemplo para o uso da base em inglês.

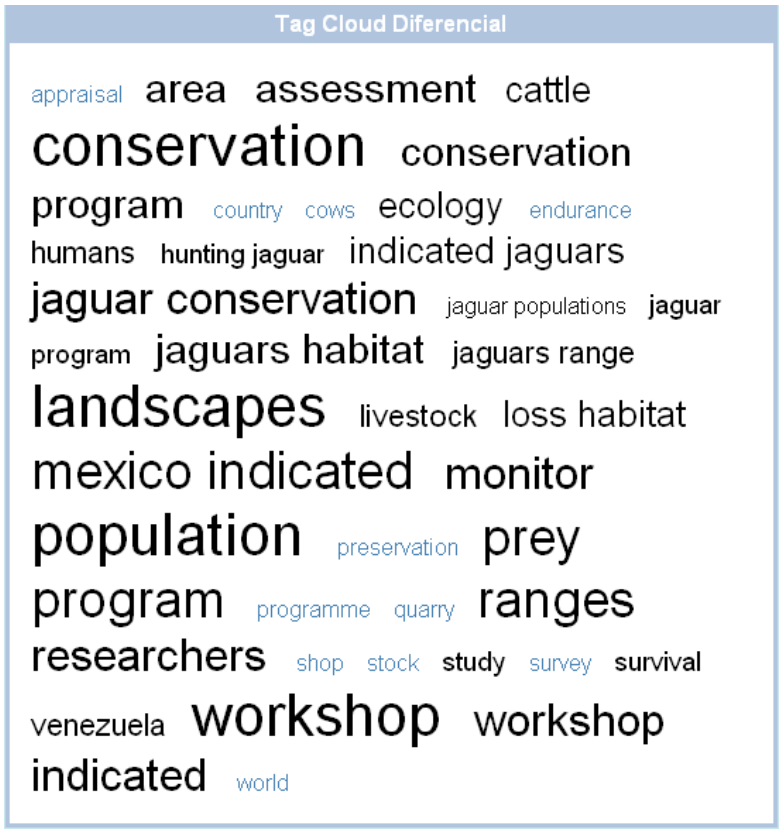


Figura 5.12: TCD de um documento referente ao animal Jaguar

Na figura 5.12 apresenta-se a TCD gerada sobre o documento[19] que trata de um programa de conservação dos jaguares. Pode-se notar *tags* na TCD que descrevem o assunto tratado no documento, tais como, "conservation", "jaguars habitat", "conservation program", "population", "preservation", etc.

[19] http://www.savethejaguar.org/

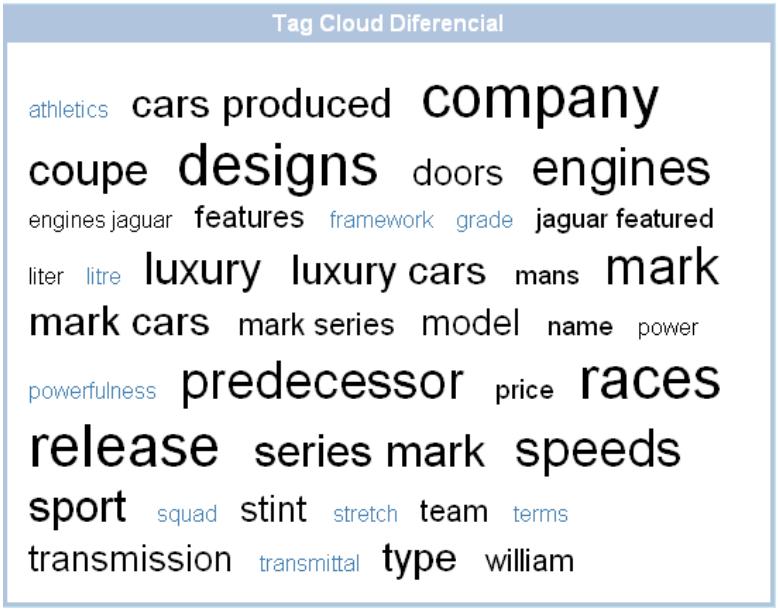


Figura 5.13: TCD de um documento referente ao fabricante de automóveis Jaguar

Na figura 5.13 é exibida a TCD para um documento[20] sobre o fabricante de carros Jaguar. Observa-se *tags* na TCD que estão relacionadas ao tema do documento, como por exemplo, “company”, “cars”, “engines jaguar”, “luxury cars”, “designs”, “races”, “sport”, etc.

[20] http://www.jaguar-automotives.com/

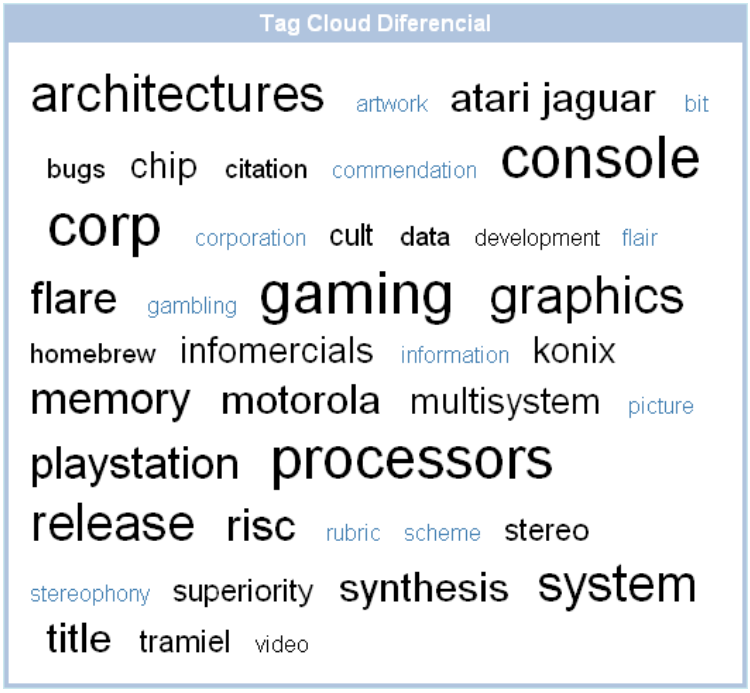


Figura 5.14: TCD de um documento referente ao console Jaguar

Na figura 5.14 apresenta-se a TCD gerada para o documento[21] que trata sobre o console de vídeo game jaguar fabricado pela Atari. Assim como nas demais TCD percebe-se *tags* que estão relacionadas ao assunto deste documento, como, “atari”, “gaming”, “graphics”, “corporation”, “console”, “system”, etc.

Em seguida, mantendo o padrão do primeiro teste foi utilizada para o teste da ferramenta uma base em português sobre Banco de Dados. Na sequência é apresentada a TCRC gerada.

[21] http://en.wikipedia.org/wiki/Atari_Jaguar

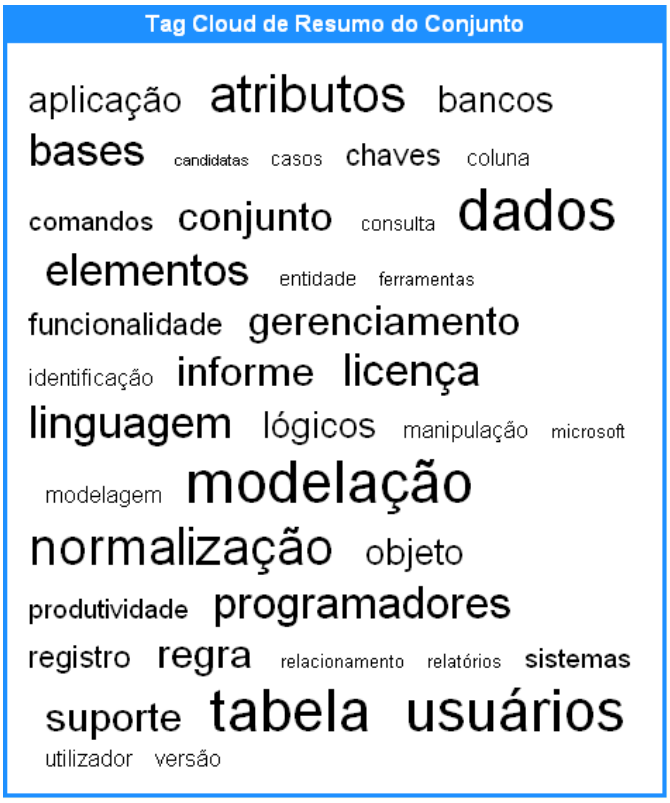


Figura 5.15: TCRC sobre a base Banco de Dados

Nesta TCRC, figura 5.15, observa-se *tags* que caracterizam o conteúdo da base utilizada neste primeiro teste, entre elas, “atributos”, “bancos”, “dados”, “gerenciamento” “modelagem”, “normalização”, “tabela”, etc.

Assim como no primeiro exemplo para esta segunda abordagem, a seguir são exibidas as 3 TCD geradas para os mesmos 3 documentos, seguindo a mesma ordem, que também foram utilizados anteriormente como exemplo para o uso da base em português.

Figura 5.16: TCD de um documento referente conceitos sobre Banco de Dados

Na figura 5.16 é apresentada a TCD gerada para um documento[22] que aborda alguns dos pontos relacionados ao conteúdo de banco de dados. Pode-se perceber na TCD *tags* que representam estes pontos, como, “armazenamento”, “base”, “bases relacionais”, “gerenciamento”, “modelo”, “transação”, etc.

[22] http://pt.wikipedia.org/wiki/Banco_de_dados

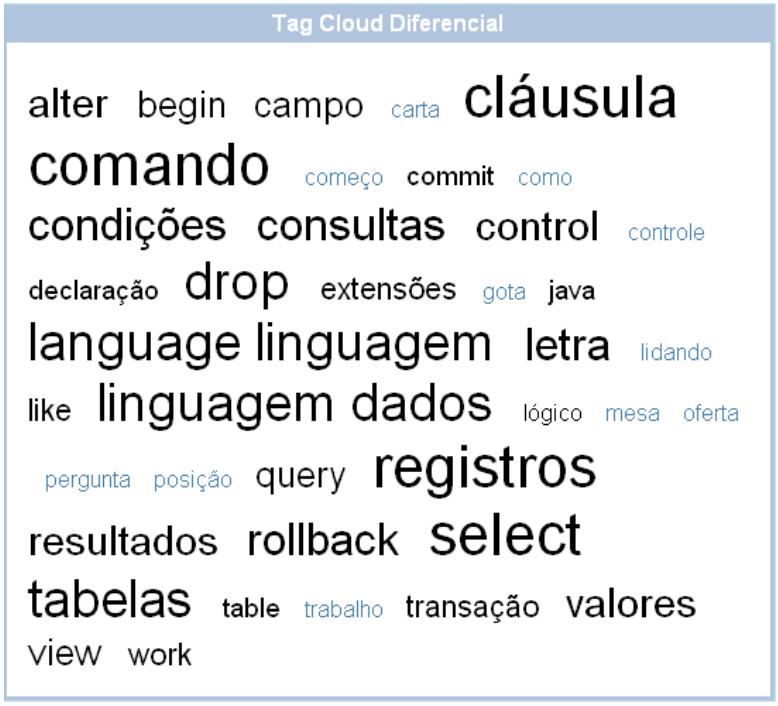


Figura 5.17: TCD de um documento referente à SQL

Na figura 5.17 exibe-se a TCD criada para o documento[23] que se refere a linguagem SQL. Nota-se, *tags* na TCD que estão relacionadas ao tema do documento, como, “cláusula”, “comando”, “condições”, “linguagem”, “registros”, “resultados”, “select”, “tabelas”, etc.

[23] http://pt.wikipedia.org/wiki/SQL

Figura 5.18: TCD de um documento referente à modelagem de dados

Na figura 5.18 mostra-se a TCD gerada sobre um documento[24] que trata do tema de modelagem de dados. Pode-se notar *tags* na TCD que estão relacionadas ao assunto abordado pelo documento, como por exemplo, “entidades”, “estrutura”, “lógico” “modelagem dados”, “negócio”, “objetos”, “relacionamentos”, “representação”, etc.

5.3 Protótipo Ferramenta

As *tag clouds* utilizadas para exemplificar os resultados foram geradas a partir de um protótipo gerado para a ferramenta proposta neste trabalho.

Para os exemplos de resultados foram utilizadas bases de teste geradas para este fim, conforme foi mencionado anteriormente. Porém, nesta ferramenta também é

[24] http://pt.wikipedia.org/wiki/Modelo_de_dados

possível que o usuário realize uma consulta na web utilizando termos em português ou inglês. Também é possível escolher, de um total de 4 possibilidades, qual abordagem ele deseja que seja utilizada para gerar as tag clouds. As 4 possibilidades são:

3. Possibilidade 1: TCRC utilizando a segunda abordagem (seção 5.2.2)
   TCD utilizando a primeira abordagem (seção 5.2.1)
4. Possibilidade 2: TCRC utilizando a segunda abordagem (seção 5.2.2)
   TCD utilizando a segunda abordagem (seção 5.2.2)
5. Possibilidade 3: TCRC utilizando a primeira abordagem (seção 5.2.1)
   TCD utilizando a primeira abordagem (seção 5.2.1)
6. Possibilidade 4: TCRC utilizando a primeira abordagem (seção 5.2.1)
   TCD utilizando a segunda abordagem (seção 5.2.2)

Na figura 5.19 exibimos uma tela da ferramenta criada.

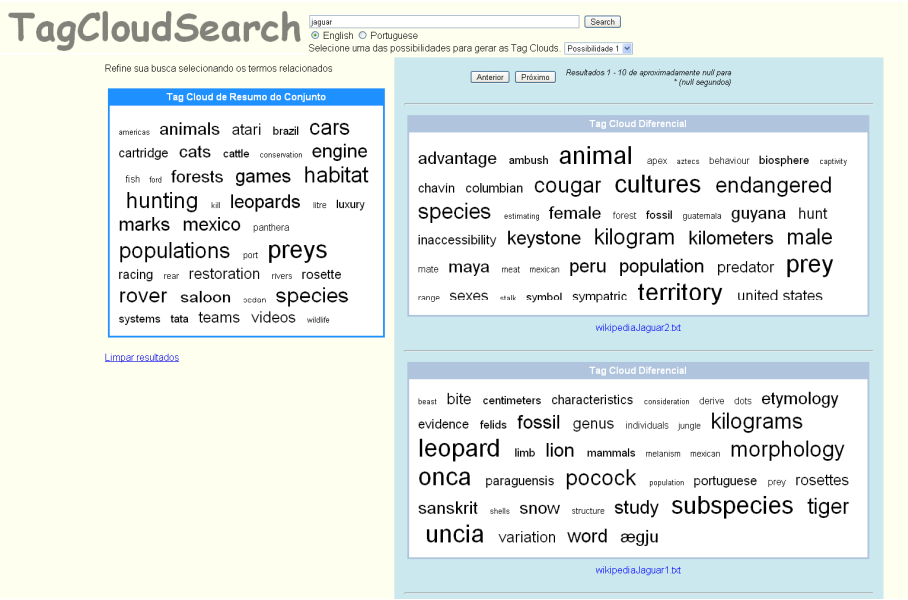


Figura 5.19: Tela da ferramenta protótipo criada

5.4 Considerações Finais

Nesta seção foram apresentados os métodos e implementações que foram utilizados neste trabalho para gerar as *tag clouds* da ferramenta que foi proposta.

Outras abordagens para a geração das *tag clouds* foram estudadas, porém, para este trabalho ficamos restritos a estas duas que foram utilizadas. Entretanto, em trabalhos futuros esperamos apresentar outras abordagens.

6. Avaliação

Foram definidas 2 formas de avaliar este trabalho. Uma avaliação numérica e outra avaliação qualitativa.

A seguir são apresentadas cada uma destas avaliações com seus respectivos resultados.

6.1 Primeira Avaliação

A proposta da primeira avaliação era fazer uma comparação das *tag clouds* obtidas pela nossa ferramenta com as *tag clouds* geradas por alguma outra ferramenta.

A ferramenta utilizada foi o *Many Eyes*[25], site no qual podemos gerar *tag clouds*, dentre outras formas de visualização, a partir de uma fonte de dados.

O objetivo desta avaliação era avaliar o processo de geração automática de *tags* para as *tag clouds* utilizados por cada uma das ferramentas comparando-o com a atribuição de *tags* feito por pessoas. Para isto foi necessário encontrar documentos que tivessem sido *tageados* manualmente. Então, foi decidido utilizar como fonte de dados para montar a base de avaliação os *bookmarks* do *delicious*[26]. Pois, para cada *bookmark* existem as *tags* relacionadas que foram definidas por pessoas.

Para poder comparar as *tag clouds* geradas pelas duas ferramentas era necessário definir um cálculo que permitisse fazer isto. Portanto, foi definido o Percentual de Coincidência de uma *tag cloud* para um documento i (P_i) da seguinte forma:

$$Pi = Ci / Ti$$

onde:

[25] http://manyeyes.alphaworks.ibm.com/

[26] http://delicious.com/

C_i = Total de *tags* coincidentes entre a *tag cloud* gerada pela ferramenta e as *tags* atribuídas pelos usuários no *delicious* para o documento *i*.

T_i = Total de *tags* existentes na *tag cloud* referente ao documento *i*.

P_i = Razão entre o total de *tags* coincidentes e o total de *tag*s presentes na *tag cloud*.

Pode-se fazer uma analogia deste cálculo com o cálculo de precisão definido na teoria de Busca e Recuperação de Informação (BRI), uma vez que ele também estabelece uma razão entre a quantidade de documentos relevantes retornados e o total de documentos recuperados por uma consulta.

Em seguida é apresentado um caso que exemplifique este cálculo:

Para um determinado *bookmark* relacionado ao documento X no *delicious* foram atribuídas as seguintes *tags*:

amistoso
brasil
futebol
mundial
seleção

Uma *tag cloud* T_1 gerada para este documento apresentou as seguintes *tags*:

brasil
campeonato
copa
equipe
futebol
seleção

Já uma outra *tag cloud* T_2 exibiu as seguintes *tags*:

africa
brasil
campeonato
cbf
classificação

eliminatória

futebol

jogo

mundial

seleção

Para T_1 o seu P_X é calculado da seguinte forma. Como T_1 apresenta 3 *tags* coincidentes com as *tags* do *bookmark* do *delicious* (brasil, futebol e seleção), seu P_X será :

$$\frac{3}{6} = 0{,}5 = 50\%$$

Logo, o P_X para T_1 é de 50%.

Já para T_2 o seu P_X é. Como T_2 apresenta 4 *tags* coincidentes com as *tags* do *bookmark* do *delicious* (brasil, futebol, mundial e seleção), seu P_X será :

$$\frac{4}{10} = 0{,}4 = 40\%$$

Logo, o P_X para T_2 é 40%.

Portanto temos para este exemplo o seguinte resultado:

	T_1	T_2
P_X	50%	40%

Tabela 6.1: Cálculo do Percentual de Coincidência para as tag clouds do exemplo

Como o P_X da *tag cloud* T_1 foi maior podemos considerar que ela apresentou um resultado mais preciso do que T_2.

A figura 6.1 exibe os valores do Percentual de Coincidência obtidos para cada documento na avaliação pelas *tag clouds* geradas pela nossa ferramenta e pelo *Many Eyes*.

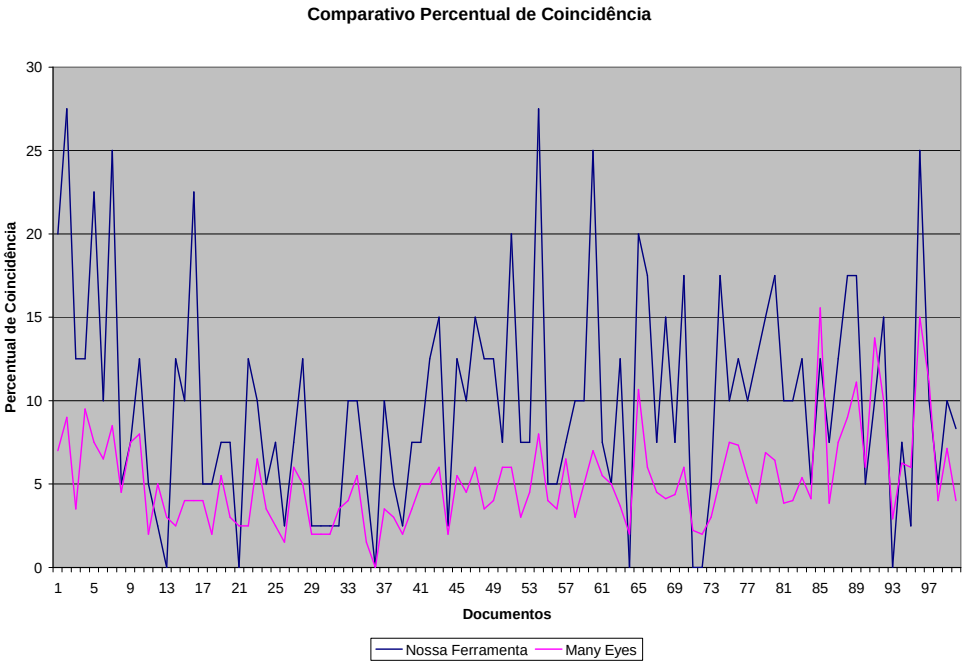


Figura 6.1: Percentual de Coincidência para as *tag clouds* geradas na avaliação

O cálculo do Percentual de Coincidência Médio (P_M) na nossa avaliação foi feito em cima de uma base que contava com 100 documentos obtidos a partir de *bookmarks* do *delicious*. O resultado obtido foi:

	Tag Clouds gerada pela nossa ferramenta	Tag Clouds gerada pelo Many Eyes
P_M	9,96%	5,19%

Tabela 6.2: Cálculo do Percentual de Coincidência Médio obtido na avaliação

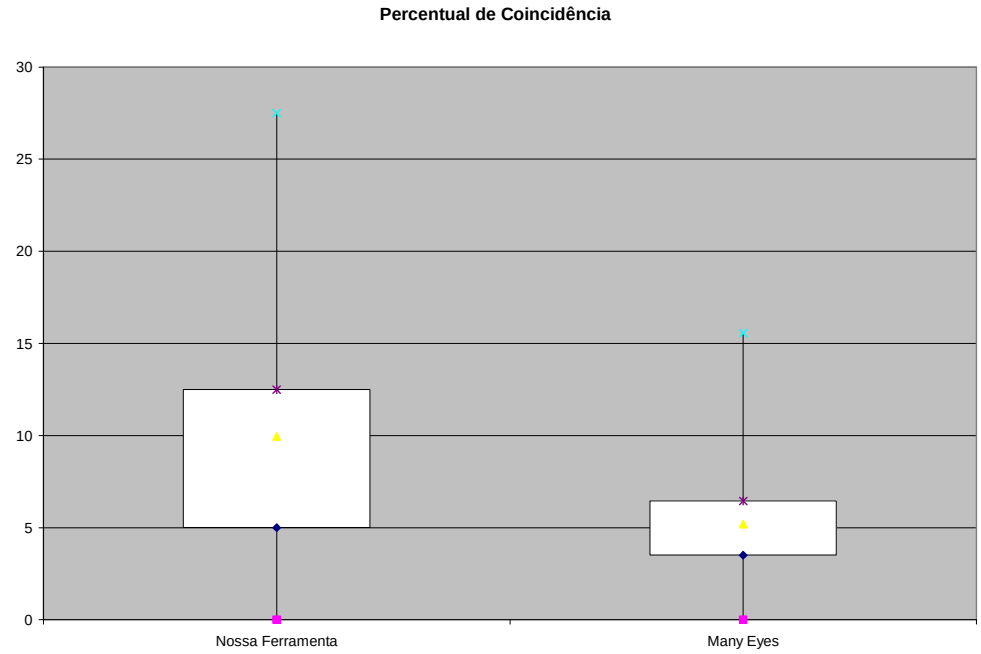


Figura 6.2: Percentual de Coincidência obtido na avaliação

Como foi feita a analogia com o cálculo de precisão, também pôde ser feito o mesmo para o cálculo de cobertura de uma *tag cloud* para um documento i (C_i).

Utilizando o exemplo dado anteriormente teríamos obtidos para as *tag clouds* T_1 e T_2 os respectivos valores para cobertura.

Para T_1 o seu C_X é calculado da seguinte forma. Como T_1 apresenta 3 *tags* coincidentes com as *tags* do *bookmark* do *delicious* (brasil, futebol e seleção), seu C_X será :

$$\frac{3}{5} = 0{,}6 = 60\%$$

Logo o C_X para T_1 é: 60%.

Já para T_2 o seu C_X é. Como T_2 apresenta 4 *tags* coincidentes com as *tags* do *bookmark* do *delicious* (brasil, futebol, mundial e seleção), seu C_X será :

$$\frac{4}{5} = 0{,}8 = 80\%$$

Logo o C_X para T_2 é: 80%.

Portanto, para este exemplo foi obtido o seguinte resultado:

	T_1	T_2
C_X	60%	80%

Tabela 6.3: Cálculo da Cobertura para as tag clouds do exemplo

A figura 6.3 mostra os valores de Cobertura alcançados para cada documento na avaliação pelas *tag clouds* geradas pela nossa ferramenta e pelo *Many Eyes*.

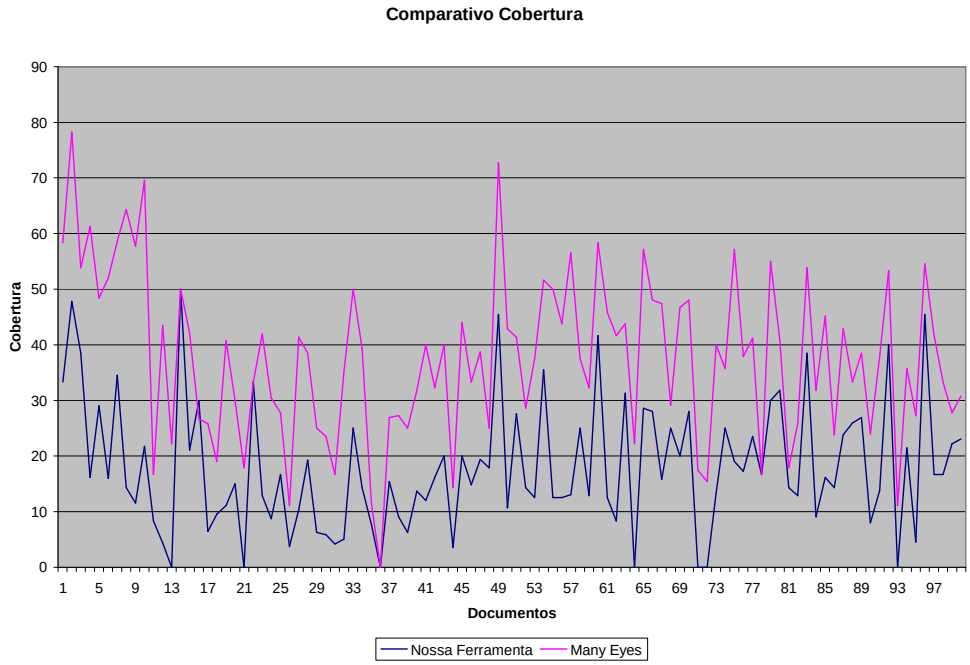


Figura 6.3: Cálculo do valor de Cobertura para as tag clouds geradas na avaliação

O resultado obtido na avaliação para os valores de Cobertura Média (C_M) é apresentado na tabela 6.4.

	Tag Clouds gerada pela nossa ferramenta	**Tag Clouds gerada pelo Many Eyes**
C_M	**17,91%**	**37,71%**

Tabela 6.4: Cálculo da Cobertura Média obtida na avaliação

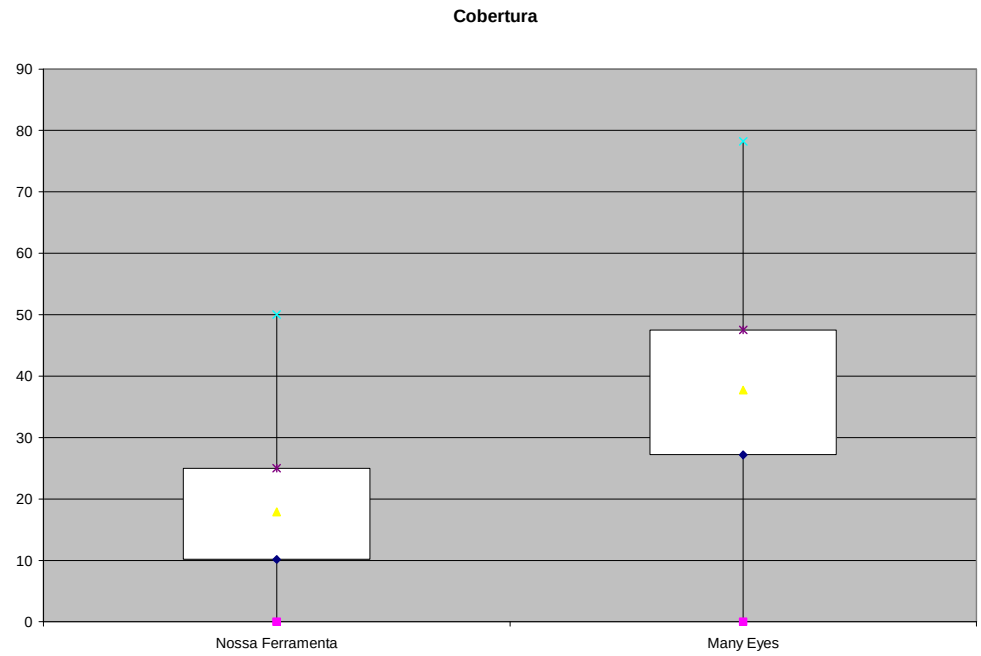


Figura 6.4: Cobertura obtida na avaliação

Analisando os resultados obtidos pode-se observar que a nossa ferramenta obteve melhores valores para o cálculo de Percentual de Coincidência e valores piores para o cálculo de Cobertura. Isto ocorre porque as *tag clouds* geradas pelo *Many Eyes* retornam um número consideravelmente maior de *tags*, observe a figura 6.5, por isso, geralmente ela apresenta uma cobertura maior do que a obtida pela nossa ferramenta.

No entanto, nossa ferramenta apresenta resultados melhores para o cálculo do Percentual de Coincidência, pois, mesmo com um número menor de *tags* conseguiu recuperar *tags* consideradas relevantes tão bem quanto à recuperação alcançada pelas *tag clouds* geradas pelo *Many Eyes*.

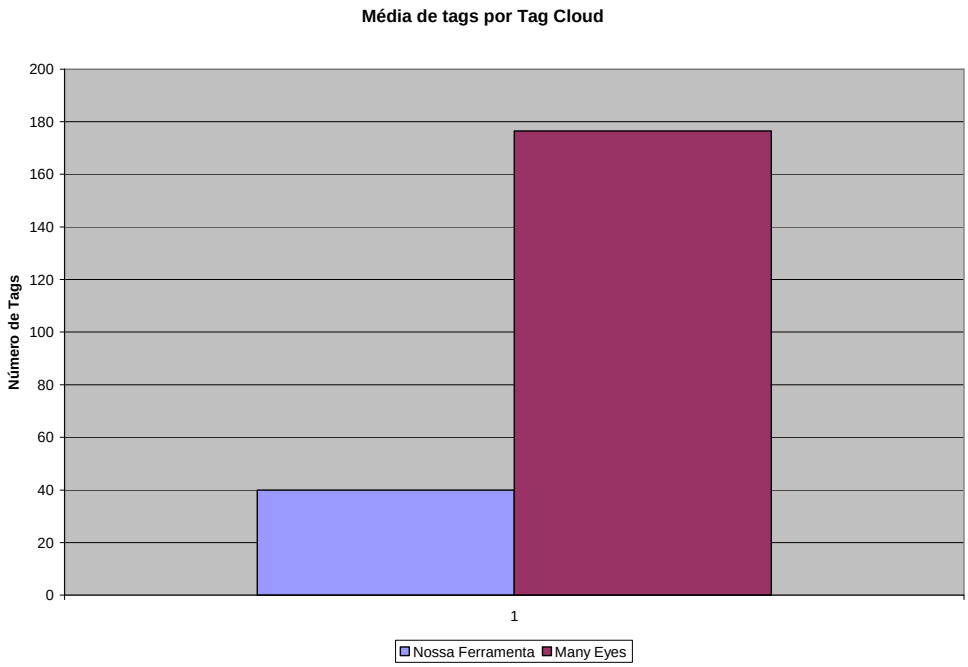


Figura 6.5: Média de Tags por Tag Cloud

6.2 Segunda Avaliação

Na segunda avaliação foi elaborado um questionário para ser respondidos por avaliadores a fim de obter mais informações sobre as *tag clouds* geradas pela nossa ferramenta.

Para essa avaliação foi montada uma nova base de documentos. Essa base possuía 20 documentos distribuídos por diferentes assuntos da seguinte forma, tabela 6.5.

Manga	**Fruta**	2
	Quadrinhos	2
	Jogador	2
	Roupa	2
Jaguar	**Animal**	2
	Carro	2
	Jogo	2
	Esporte	2
Inteligência Computacional		2
Banco de Dados		2
Total		20

Tabela 6.5: Distribuição dos assuntos da base

Nesta segunda avaliação participaram 16 avaliadores, todos com formação superior e diferentes níveis de conhecimento sobre *tag clouds.*

O questionário possuía 18 questões que estavam dividas em 4 grupos de tipo de questões, cada grupo com um objetivo de avaliar determinado fator.

O primeiro grupo de questões do questionário pretendia avaliar a percepção do avaliador em relação à representatividade das *tag clouds* geradas para cada documento respectivamente. Para avaliar a representatividade os avaliadores deveriam verificar se as *tags* apresentadas na *tag cloud* tinham alguma relação com o assunto do documento que era citado na questão.

Os avaliadores poderiam escolher 1 entre 5 níveis para responder se concordavam ou não com as *tag* presentes na *tag clouds*. Os níveis existentes para a avaliação eram (Concordo totalmente; Concordo parcialmente; Não concordo nem discordo; Discordo parcialmente; Discordo totalmente).

Foram apresentadas aos avaliadores 6 questões, cada uma abordando um assunto diferente. Na tabela 6.6 é apresentada quantos votos em média cada nível recebeu.

Nível	Média
Concordo totalmente	2,83
Concordo parcialmente	7,83
Nem concordo nem discordo	2,33
Discordo parcialmente	2,5
Discordo totalmente	0,5

Tabela 6.6: Média de votos recebidos por cada nível

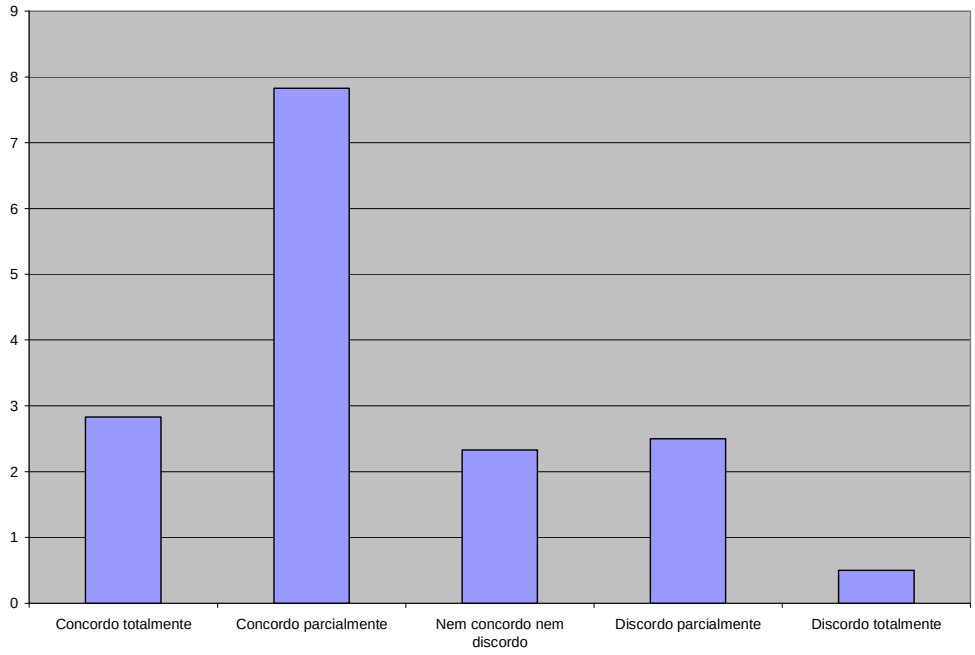


Figura 6.6: Média de votos recebidos por cada nível

Com este primeiro resultado percebemos que a maioria dos avaliadores considerou que as *tags* apresentadas nas *tag clouds* conseguiam produzir uma boa representação do assunto tratado no documento relacionado a respectiva *tag cloud*.

No segundo grupo de questões, também pretendíamos avaliar a percepção dos avaliadores quanto à representatividade das *tag clouds*, assim como no primeiro grupo. No entanto, desta vez queríamos avaliar as suas percepções em relação à representatividade das *tag clouds* de resumo do conjunto.

O método de avaliação também era semelhante ao do grupo de questões anterior. Cada avaliador deveria atribuir um nível de avaliação a *tag cloud* de resumo do

conjunto apresentada em cada questão. Na questão era mencionada com qual assunto aquela *tag cloud* havia sido gerada.

Na tabela 6.7 é apresentado a média de votos recebido por cada nível.

Nível	Média
Concordo totalmente	6,5
Concordo parcialmente	6
Nem concordo nem discordo	2,5
Discordo parcialmente	0,5
Discordo totalmente	0,5

Tabela 6.7: Média de votos recebidos por cada nível

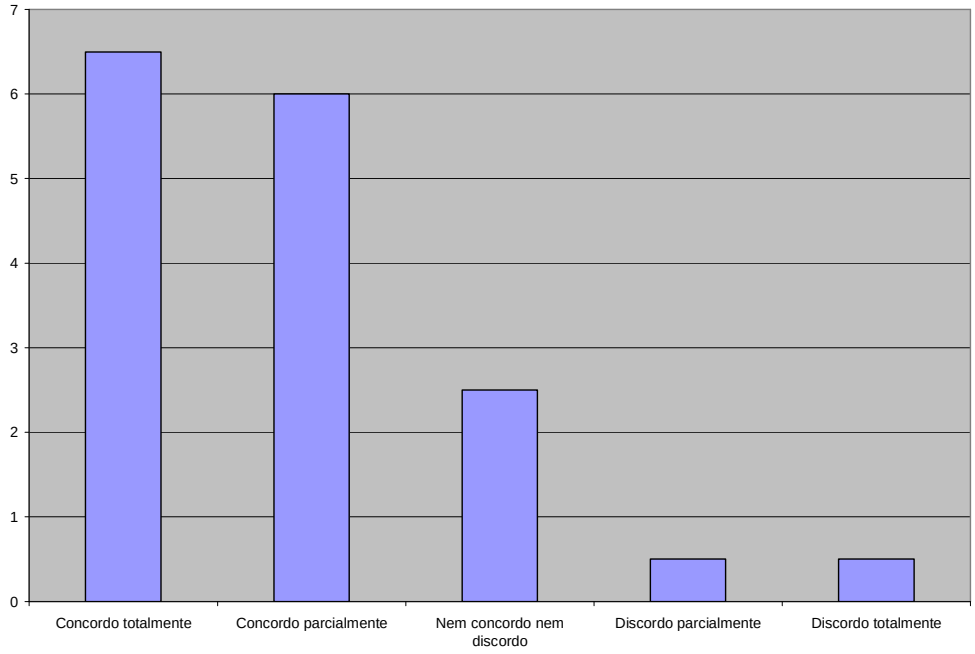


Figura 6.7: Média de votos recebidos por cada nível

No terceiro grupo de questões, foram apresentadas 3 questões. Em cada uma delas era exibida uma *tag cloud* diferencial e era solicitado para que o avaliador identificasse entre os 7 textos do grupo de questões anterior, de qual deles havia sido gerado aquela *tag cloud* diferencial.

De forma surpreendente todos os avaliadores acertaram na correlação das 3 questões, tabela 6.8.

Isto nos dá um forte indício de que as *tag clouds* diferenciais conseguiram cumprir o objetivo para o qual elas foram criadas. Ou seja, as *tag clouds* diferenciais foram capaz de exibir *tags* que exibissem propriedades e características do texto associado a ela.

Correlação	Quantidade
Correta	100%
Errada	0%

Tabela 6.8: Correlação entre *tag clouds* diferenciais e seus respectivos textos

No quarto e último grupo de questões, existiam sete questões e em cada uma delas foi solicitado para que os avaliadores atribuíssem *tags* a cada um dos textos exibidos nestas questões. Nosso objetivo nessa questão era comparar e avaliar o desempenho de atribuição de *tags* a partir da nossa ferramenta em relação à atribuição de *tags* feitas por pessoas.

Vale ressaltar neste caso que a atribuição de *tags* feitas de forma automática tem muito mais limitações do que a atribuição feita por pessoas. Por exemplo, pessoas podem atribuir *tags* a partir de palavras não contidas no texto, podem fazer o relacionamento com sinônimos mais facilmente, não fica restrito ao idioma do texto entre outras.

Para realizar esta comparação foi realizado o seguinte procedimento: listamos todas as *tags* atribuídas pelos avaliadores para cada texto e computamos por quantos avaliadores cada uma delas foi atribuída.

Após fazer isto para cada um dos textos, identificamos para cada texto, quantas das *tags* atribuídas pelos usuários eram coincidentes com as *tags* presentes nas *tag clouds* geradas respectivamente para o mesmo texto por nossa ferramenta.

Então, com estes dois valores obtidos: o número total de *tags* atribuídas ao texto e o número de *tags* coincidentes, calculamos o aproveitamento percentual para cada texto fazendo a divisão do segundo valor pelo primeiro. Por exemplo, supomos para que

um documento fosse atribuído ao todo 10 *tags*. Destas 10 *tags* 3 fossem coincidentes com *tags* presentes na *tag cloud* para esse mesmo texto. Portanto, o aproveitamento seria de:

$$\frac{3}{10} = 0{,}3 = 30\%$$

Após calcular este percentual para cada um dos textos foi calculada a média para todos os textos presentes nestas questões. O resultado obtido é apresentado na tabela 6.9.

	Questões							
	1	2	3	4	5	6	7	Percentual Médio
Média	0,19	0,24	0,05	0,33	0,13	0,49	0,32	0,25 (25%)

Tabela 6.9: Percentual Médio de *tags* coincidentes

Portanto, o resultado obtido foi de 25% de aproveitamento. Ou seja, a cada 4 *tags* que eram atribuídas pelos avaliadores, pelo menos uma era coincidente com uma das *tags* presentes nas *tag clouds* geradas pela nossa ferramenta.

Considerando as limitações apresentadas anteriormente pela atribuição de *tags* por pessoas em relação a atribuição de *tags* feitas automaticamente, esse resultado alcançado pode ser considerado bastante razoável.

Possivelmente, com algumas melhorias no processo de geração de *tag clouds* esse resultado pode ser melhorado.

6.3 Considerações Finais

Nesta seção foi apresentada a avaliação realizada neste trabalho. Foram realizados dois tipos de avaliação.

A primeira delas com o objetivo de comparar o método de geração de *tag clouds* com outro método já disponível na internet (*Many Eyes*). De acordo com as medidas utilizadas para comparar as duas abordagens, nossa ferramenta apresentou melhores

resultados, pois, conseguiu uma recuperação de *tags* de forma mais eficiente do que a alcançada pela ferramenta do *Many Eyes*.

Na segunda avaliação era pretendido conhecer o sentimento das pessoas em relação às *tag clouds* geradas em relação a distintos aspectos, como utilidade, qualidade, eficiência, etc. E de acordo com a análise dos resultados obtidos, as *tag clouds* geradas conseguiram exercer o papel para os quais elas foram propostas.

Portanto, analisando de uma forma geral os resultados alcançados pode-se considerar que os resultados alcançados nas avaliações foram bastante satisfatórios.

7. Conclusão

Com o aumento da quantidade de ferramentas ligadas a web social, a ação de *tagging* vem sendo cada vez mais utilizada. Com isto surgiu a necessidade da criação de uma nova forma para apresentar estas *tags*. Disso veio o surgimento e a utilização das *tag clouds* para preencher esta lacuna.

Neste trabalho foi proposta uma nova possibilidade para a utilização das *tag clouds*: usá-las conjuntamente com uma ferramenta de busca.

Esta idéia surgiu a partir da identificação de duas características que poderiam ser adicionadas às ferramentas de busca atuais existentes. Primeiro, a possibilidade de o usuário ter uma visão geral dos resultados retornados pela consulta e segundo, permitir que o usuário consiga distinguir características e conceitos de cada resultado que o diferencie dos demais.

Para viabilizar isto, primeiramente, foi definido conceitualmente o que seriam estes dois tipos de *tag clouds*.

1. Uma *tag cloud* para resumir os assuntos e conceitos abordados nos documentos retornados na busca. Chamamos esta *tag cloud* de ***"Tag Cloud* de Resumo do Conjunto"**.
2. Uma *tag cloud* para permitir a visualização das características que distinguem um determinado documento em relação aos demais documentos retornados na busca. Chamamos esta *tag cloud* de ***"Tag Cloud* Diferencial"**.

Para permitir a criação destes dois tipos de *tag clouds*, também foi estabelecido neste trabalho um modelo formal para a construção de *tag clouds*. Este modelo foi criado com o objetivo de que também possa ser utilizado por outras ferramentas que façam o uso de *tag clouds*.

Com os resultados apresentados e considerando a dificuldade de capturar conceitos e características associados aos documentos, podemos considerar que os objetivos deste trabalho foram atingidos.

7.1 Trabalhos Futuros

Como trabalhos futuros foram identificados alguns pontos que poderiam ser aprimorados em relação ao trabalho atual, bem como, outros que não foram contemplados por este.

A utilização de outras técnicas de modelagem de tópicos para a identificação de conceitos presentes nos documentos.

Adaptação da ferramenta proposta para outras línguas além do inglês e português.

Permitir que as *tag clouds* exibissem *tags* independente de qual é o idioma do documento correspondente. Por exemplo, em uma *tag cloud* gerada a partir de um documento em inglês exibir *tags* em português ou qualquer outro idioma que o usuário deseja.

Um dos aspectos que também consideramos que poderia ser melhorado durante a construção da ferramenta foi o fator desempenho. Todo o processamento dos textos era relativamente custoso. E como o uso desta ferramenta seria no ambiente web, esse aspecto deveria ser estudado com bastante atenção para que os fatores desempenho e tempo de resposta, não inviabilizassem o uso desta ferramenta em um ambiente que exige respostas em tempos relativamente curtos como a web.

Outra possibilidade para trabalhos futuros seria a execução de outras formas de avaliação do fator qualidade associado às *tag clouds* e da utilidade do seu uso no contexto sugerido neste trabalho.

8. Referências Bibliográficas

Anderson, J. R., 1990, *The adaptive character of thought*. 1 ed. Psychology Press.

Bateman, S., Gutwin, C., Nacenta, M., 2008. “Seeing things in the clouds: the effect of visual features on tag cloud selections”. In: *Proceedings of the Nineteenth ACM Conference on Hypertext and Hypermedia, HT '08*, pp.193-202, Pittsburgh, PA, USA, June 19 - 21, 2008.

Berners-Lee, T., Fielding, R. Masinter, L., 2005, *Uniform Resource Identifier: Generic Syntax, RFC 3986*, January 2005.

Bielenberg, K., Zacher, M., 2006, *Groups in Social Software: Utilizing Tagging to Integrate Individual Contexts for Social Navigation*. M.Sc. dissertation, Unisersitat Bremen.

Croft, B., Metzler, D., Strohman T., 2009, *Search Engines: Information Retrieval in Practice*. 1 ed. Addison Wesley.

Dasgupta A., Drineas P., Harb B., *et al.*, 2007, “Feature Selection Methods for Text Classification”. In: *Proceedings of the 13th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining*, KDD '07, pp. 230-239, San Jose, California, USA, August 12 - 15, 2007.

Deerwester, S., Dumais, S. T., Furnas, G. W., *et al.*, 1990, "Indexing by Latent Semantic Analysis". *Journal of the American Society for Information Science*, pp. 391–407.

Dubinko, M., Kumar, R., Magnani, J., *et al.*, 2006, “Visualizing Tags over Time”. In: *Proceedings of the 15th international Conference on World Wide Web, WWW '06*, pp.193-202, Edinburgh, Scotland, May 23 - 26, 2006.

Dumais, S., 1995, “What You Get Is What You Want: Combining Evidence for Effective Information Filtering”. In: *Proceedings of the Eighteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 369.

Fellbaum, C., 1998, *WordNet: An Electornic Lexical Database*. 1 ed. MIT Press.

Friedman, V., Lennartz, S., 2007, Tag Clouds Gallery: Examples And Good Practices. Disponível em: <http://www.smashingmagazine.com/2007/11/07/tag-clouds-gallery-examples-and-good-practices/> . Acesso em 10/09/2010.

Griffiths, T., Steyvers, M., "Probabilist Topic Models". In: *Handbook of Latent Semantic Analysis*, 1ed, chapter 21, Psychology Press, 2007.

Halvey, M. J., Keane, M. T., 2007, "An assessment of tag presentation techniques". In *Proceedings of the 16th international Conference on World Wide Web*, WWW '07, pp. 1313-1314, Banff, Alberta, Canada, May 08 - 12, 2007).

Hassan-Montero Y. and Herrero-Solana V., 2006, "Improving Tag-Clouds as Visual Information Retrieval Interfaces". In: *Proceedings of the 1st International Conference on Multidisciplinary Information Sciences and Technologies InSCiT*.

Hearst, M. A., Rosner, D., 2008b, "Tag Clouds: Data Analysis Tool or Social Signaller?". In: *Proceedings of the Proceedings of the 41st Annual Hawaii international Conference on System Sciences HICSS*, Washington, DC, January 07 - 10, 2008.

Heymann, P., Ramage, D., Garcia-Molina, H. 2008. "Social tag prediction". In: *Proceedings of the 31st Annual international ACM SIGIR Conference on Research and Development in information Retrieval, SIGIR '08*, pp.531-538, Singapore, Singapore, July 20 - 24, 2008.

Kuo B.Y., Hentrich T., Good B. M., *et al.*, 2007, "Tag clouds for summarizing web search results". In: *Proceedings of the 16th international Conference on World Wide Web WWW '07*, pp. 1203-1204, New York, NY, 2007.

Lamantia J. 2006. Tag Clouds: Navigation for Landscapes of Meaning. Disponível em: <http://www.joelamantia.com/blog/archives/ideas/tag_clouds_navigation_for_landscapes_of_meaning.html>. Acesso em: 09/09/2010.

Lamantia, J., 2007, Text Clouds: A New Form of Tag Cloud?. Disponível em: <http://www.joelamantia.com/blog/archives/tag_clouds/text_clouds_a_new_form_of_tag_cloud.html>. Acesso em 10/09/2010.

Landauer, T. K., Foltz, P. W., Laham D., 1998, “An Introduction to Latent Semantic Analysis”. In: *Discourse Processes*, pp. 259-284.

Laurence, S. Margolis, E., 1999, *Concepts and Cognitive Science.* MIT Press.

Manning, C., Raghavan, P., Schtze, H., 2008, *Introduction to Information Retrieval*. Cambridge University Press.

Manola, F., Miller, E., 2004, RDF Primer, W3 Recommendation. Disponível em: < http://www.w3.org/TR/rdf-primer/>. Acesso em 10/09/2010.

Marinchev I., 2006, “Practical Semantic Web – Tagging and Tag Clouds”, *Journal Cybernetics and Information Technologies*, v. 6, n. 3, pp. 33 – 39.

Penev, A., Wong, R. K., 2008, “Finding similar pages in a social tagging repository”. In: *Proceeding of the 17th international Conference on World Wide Web, WWW '08*, pp.1091-1092, Beijing, China, April 21 - 25, 2008.

Porter M.F., 1980, *An algorithm for suffix stripping*. Program, 14, pp. 130–137.

Rivadeneira A. W., Gruen D. M., Muller M. J., *et al.*, 2007, “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*, pp 995-998, New York, NY, 2007.

Rodgers, J. L., Nicewander, A. W., 1988, “Thirteen Ways to Look at the Correlation Coefficient”. In: *The American Statistician*, pp.59-66.

Salton, G., 1968, *Automatic Information Organization and Retrieval*. McGraw Hill Text.

Shaw, B., 2005, Utilizing Folksonomy : Similarity Metadata from the Del.icio.us System. Disponível em: <http://www.metablake.com/webfolk/web-project.pdf>. Acesso em 10/09/2010.

Silverstein, C., Marais, H., Henzinger, M., *et al.*, 1999, “Analysis of a very large web search engine query log”. *SIGIR Fórum*, pp. 6-12, September, 1999.

Song, Y., Zhuang, Z., Li, H., *et al.*. 2008. “Real-time automatic tag recommendation”. In: *Proceedings of the 31st Annual international ACM SIGIR Conference on Research and Development in information Retrieval, SIGIR '08*, pp. 515-522, Singapore, Singapore, July 20 - 24, 2008.

Viégas, F. B., Wattenberg, M., 2008. “TIMELINES - Tag clouds and the case for vernacular visualization”. In: *Interactions*, pp.49-52, July, 2008.

Watters D., 2009, Meaningful Clouds: Towards a novel interface for document visualization. Disponível em: <http://danwatters.com/documents/CloudMine_dwatters.pdf>. Acesso em 09/09/2010.

Wei, X., 2007, *Topic Models in Information Retrieval*. Ph.D. dissertation, University of Massachusetts Amherst,.

Weiss S., Indurkia N., Zhang T., *et al.*, 2005, Text Mining Predictive Methods for Analyzing Unstructured Information. Springer.

Xexéo G., Morgado F., Fiuza P., 2009a. “Automatically Generated Tag Clouds”. *In Proceedings of the XXIV Simpósio Brasileiro de Banco de Dados, SBBD 2009*, pp. 136-150, Fortaleza, October, 2009.

Xexéo, G., Morgado, F., and Fiuza, P. 2009b. “Differential Tag Clouds: Highlighting Particular Features in Documents”. In *Proceedings of the 2009 IEEE/WIC/ACM international Joint Conference on Web intelligence and intelligent Agent Technology - Volume 03*, WI-IAT.2009, pp. 129-132, September 15 - 18, 2009.

Anexo I – Questionário Avaliação

ORIENTAÇÃO

O questionário a seguir apresenta 18 questões de 4 tipos distintos divididas em blocos. Cada bloco será precedido de sua respectiva explicação.

Definições:

***1) Tags* são rótulos que de alguma forma caracterizam um objeto ou alguma propriedade/característica dele. As *tags* podem ser definidas a partir do que você considere mais relevante sobre o assunto apresentado no texto. A *tag* pode ser ou não uma palavra existente no texto.**

***2)* Uma *Tag Cloud* é uma representação visual de um conjunto de palavras (*tags*) de tal forma que possam transmitir informação e significado através de propriedades como tamanho, cor, estilo, etc.**

***3)* Uma *Tag Cloud Diferencial* tem por objetivo apresentar as principais características e particularidades de um documento que permitem diferenciá-lo dentro de uma coleção de documentos.**

4) Uma *Tag Cloud de Resumo do Conjunto* tem por objetivo apresentar um resumo dos assuntos tratados pelos documentos de um conjunto específico.

Questões 1 a 7: É apresentado um texto e você deve atribuir ao texto algumas *tags* que considere relevantes.

Exemplo:

Os organizadores da Copa do Mundo levaram nesta quinta-feira a taça da competição a Nelson Mandela, seis anos depois dele ser fotografado com o troféu no dia em que a África do Sul foi escolhida para sediar o Mundial em 2010. O ex-presidente sul-africano, de 91 anos, se encontrou com Jerome Valcke, secretário-geral da Fifa, e Danny Jordaan, chefe do comitê organizador local, que levaram o troféu de 18 quilates até o prédio da Fundação Nelson Mandela em Johannesburgo.

Valcke disse que Mandela foi um dos arquitetos da Copa do Mundo de 2010. "Para nós, não havia maneira do troféu chegar ao país e não ser levado primeiro para Mandela". O troféu da Copa do Mundo começa uma turnê nacional na África do Sul na sexta-feira, a partir da Cidade do Cabo.

Jordaan disse que foi "um momento emocionante e alegre" trazer o troféu de volta para Mandela, que é carinhosamente conhecido como "Madiba" pelos sul-africanos, um título tradicional adotado pela membros de seu clã.

"Foi tão maravilhoso ver Madiba, em Zurique, no dia que a África do Sul ganhou o direito de sediar a Copa do Mundo de 2010", disse Jordaan. "Tão feliz com lágrimas de alegria rolando pelo seu rosto. Portanto, é um momento muito emocionante e alegre trazer de volta o troféu para ele, um símbolo do futebol mundial com um símbolo mundial da humanidade".

Mandela tem feito raras aparições públicas e é incerto se o cidadão mais famoso da África do Sul estará presente na cerimônia de abertura da Copa do Mundo, no Estádio Soccer City, em 11 de junho, embora o presidente da Fifa, Joseph Blatter, disse esperar que o ex-presidente compareça.

(...)

Tags: futebol, copa, mundial, mandela, taça

1. Quais tags você atribuiria a este documento?

A manga é considerada uma das frutas mais delicadas e a rainha das espécies tropicais, sendo até citada nas escrituras budistas. É uma fruta suculenta, de sabor muito exótico. Está presente em diversas regiões do mundo. A manga é fruto da mangueira que se desenvolve em condições de clima subtropical.

Planta originária do sul da Ásia, atualmente é cultivada em praticamente todos os países de clima tropical e subtropical. A mangueira pode atingir entre 35 e 40 m de altura. A fruta tem tamanho e formato variado e sua coloração pode variar em amarelo, laranja e vermelho. Quando madura, sua cor tende a ser amarelada, porém pode ocorrer da fruta estar madura, mas com coloração verde. A fruta pode ser degustada in natura, pois é rica em vitaminas, minerais e antioxidantes. Por conter uma grande quantidade de ferro é bastante indicada para tratamentos de anemia. É uma das frutas mais consumidas em todo o mundo. É contabilizado hoje um número entre 500 e 1000 variedades existentes.

Os estados de São Paulo e Minas Gerais são os maiores produtores de manga do país, juntos alcançam cerca de 50% da área plantada e 25% do total da produção, logo em seguida vêm os estados da Bahia, Pernambuco, Piauí e Ceará. A fruta é comercializada tanto no território nacional como também em outros países.

Quem já não ouviu dizer: "manga com leite não pode". Pode sim, é tudo história. Assim, vale a pena aproveitar a safra que vai de novembro até janeiro e se esbaldar com a fruta, seja batida com leite, em saladas, molhos ou doces. "A combinação leite com manga não faz mal, mas existem outras combinações que ficam melhor com a manga", afirma a nutricionista Milena Teixeira.

Mas é preciso certa atenção ao comer manga à noite. "Não faz mal pra saúde, mas algumas pessoas podem ficar com uma indisposição, com dor no estômago", explica.

A mangueira, originária da Índia, é uma árvore tropical que pertence à mesma família do cajueiro. O fruto varia muito em tamanho e cor. As mangas menores são do tamanho de uma nêspera, enquanto as maiores podem pesar até 2 kg. Quanto à forma, existem as redondas, ovais, alongadas e finas, do formato de um coração e até mesmo de um rim.

Em relação à cor, podem ter casca bem verde, amarela ou vermelha, de acordo com a variedade. A polpa da manga é suculenta, com sabor bem característico, algumas vezes fibrosa e de cor que varia do amarelo-claro ao alaranjado-escuro.

2. Quais tags você atribuiria a este documento?

O mangá é um gênero da literatura japonesa e é caracterizado por seus quadrinhos. Os mangás têm suas raízes no período Nara (século VIII d.C.), com o aparecimento dos primeiros rolos de pinturas japonesas: os *emakimono*. Eles associavam pinturas e textos que juntos contavam uma história à medida que eram desenrolados. O primeiro desses *emakimono*, o *Ingá Kyô*, é a cópia de uma obra chinesa e separa nitidamente o texto da pintura.

A partir da metade do século XII, surgem os primeiros *emakimono* com estilo japonês. O *Genji Monogatari Emaki* é o exemplar de *emakimono* mais antigo conservado, sendo o mais famoso o *Chojugiga*, atribuído ao bonzo Kakuyu Toba e preservado no templo de Kozangi em Kyoto. Nesses últimos surgem, diversas vezes, textos explicativos após longas cenas de pintura. Essa prevalência da imagem assegurando sozinha a narração é hoje uma das características mais importantes dos mangás.

No período Edo, em que os rolos são substituídos por livros, as estampas eram inicialmente destinadas à ilustração de romances e poesias, mas rapidamente surgem livros *para ver* em oposição aos livros *para ler*, antes do nascimento da estampa independente com uma única ilustração: o ukiyo-e no século XVI. É, aliás, Katsushika Hokusai o precursor da estampa de paisagens, nomeando suas célebres caricaturas publicadas de 1814 à 1834 em Nagoya, cria a palavra mangá — significando "desenhos irresponsáveis".

Os mangás não tinham, no entanto, sua forma atual, que surge no início do século XX sob influência de revistas comerciais ocidentais provenientes dos Estados Unidos e Europa. Tanto que chegaram a ser conhecidos como *Ponchie* (abreviação de *Punch-picture*) como a revista britânica, origem do nome, Punch Magazine (Revista

Punch), os jornais traziam humor e sátiras sociais e políticas em curtas tiras de um ou quatro quadros.

Diversas séries comparáveis as de além-mar surgem nos jornais japoneses: Norakuro Joutouhei (Primeiro Soldado Norakuro) uma série antimilitarista de Tagawa Suiho, e Boken Dankichi (As aventuras de Dankichi) de Shimada Keizo são as mais populares até a metade dos anos quarenta, quando toda a imprensa foi submetida à censura do governo, assim como todas as atividades culturais e artísticas. Entretanto, o governo japonês não hesitou em utilizar os quadrinhos para fins de propaganda.

Sob ocupação americana após a Segunda Guerra Mundial, os mangakas, como os desenhistas são conhecidos, sofrem grande influência das histórias em quadrinhos ocidentais da época, traduzidas e difundidas em grande quantidade na imprensa cotidiana.

3. Quais tags você atribuiria a este documento?

Na Ilha do Retiro, estádio e toca do leão Sport Club do Recife, quem olhasse aquele rapaz esguio, sério e de rosto marcado pela varíola, via um belo projeto de goleiro. Isso em 1954, quando ele chegou à equipe juvenil do Sport aos 17 anos e com 1,86 m de altura. Nisso, Lula Carlos – repórter arguto e cronista antenado – quis saber mais. E o guardião tímido, adolescente sem traquejo verbal, disse ser Haílton Corrêa Arruda, ou Manga, nascido em 26 de abril de 1937, no Pina – bairro da capital pernambucana onde havia vindo ao mundo um outro craque: Ademir Menezes, o Queixada.

Nesse 54, sem tomar gol, Manga foi campeão juvenil invicto ao lado de Almir Pernambuquinho, meia de muita bola no pé e nenhum juízo no quengo. Mas o dia do jovem goleiro veio numa excursão à Europa, em 56, com Oswaldo Baliza contundido em Lisboa e ele assumindo o arco do rubro-negro recifense, inclusive diante do Real Madrid, no estádio Santiago Bernabeu. Ano seguinte, com físico privilegiado, coragem, elasticidade, mãos enormes – que sofreriam inúmeras fraturas –, firmeza no jogo aéreo, ótima colocação e reflexo, Manga se notabilizou no Brasil. Porém, só foi para o Botafogo carioca viver o auge da carreira no início de 1959, tendo antes sido campeão estadual pernambucano pelo Sport, atuando com velhas raposas do futebol, como o craque carioca Mirim e o alagoano Tomires.

No Rio, criaram até prêmio para quem fizesse gol nele. E Nilo, do América, levou um aparelho de televisão em 25 de abril de 1959, véspera do aniversário do arqueiro recifense. Em 60, Manga venceu o torneio internacional de Bogotá. E em 62 os bicampeonatos dos torneio início e certame cariocas, ganhando ainda o 6º pentagonal do México e o torneio Rio-São Paulo. Sua performance fez Nílton Santos, o Enciclopédia, confessar: “Foi o melhor goleiro que passou pelo Botafogo desde que me conheço por

gente”. E é possível que Garrincha, Didi e Amarildo – os outros artistas do alvinegro – pensassem o mesmo do guarda-redes, a quem chamavam de Manguinha.

Mas, viciado e perdedor no baralho, Manga tomava empréstimos para cobrir dívida da jogatina. E depois pedia mais dinheiro ao clube. Para o jornalista João Saldanha, tal penúria expunha seu frágil caráter – e, adiante, isso seria o estopim da desavença entre eles. No entanto, em campo, o goleiro erguia troféus. Em 1963, foi tricampeão do torneio início e venceu o 4º torneio de Paris. Ano seguinte – com a liberdade no Brasil indo a pique em um reles golpe militar –, ele pôs o time na conquista destes títulos: torneio Rio-São Paulo, jubileu de ouro da federação de futebol da Bolívia, quadrangular ibero-americano de Buenos Aires e taça Magalhães Pinto, em Belo Horizonte. (...)

4. Quais tags você atribuiria a este documento?

Jaguars are the largest of South America's big cats. They once roamed from the southern tip of that continent north to the region surrounding the U.S.-Mexico border. Today significant numbers of jaguars are found only in remote regions of South and Central America—particularly in the Amazon basin.

These beautiful and powerful beasts were prominent in ancient Native American cultures. In some traditions the Jaguar God of the Night was the formidable lord of the underworld. The name jaguar is derived from the Native American word *yaguar*, which means "he who kills with one leap."

Unlike many other cats, jaguars do not avoid water; in fact, they are quite good swimmers. Rivers provide prey in the form of fish, turtles, or caimans—small, alligatorlike animals. Jaguars also eat larger animals such as deer, peccaries, capybaras, and tapirs. They sometimes climb trees to prepare an ambush, killing their prey with one powerful bite.

Most jaguars are tan or orange with distinctive black spots, dubbed "rosettes" because they are shaped like roses. Some jaguars are so dark they appear to be spotless, though their markings can be seen on closer inspection.

Jaguars live alone and define territories of many square miles by marking with their waste or clawing trees.

Females have litters of one to four cubs, which are blind and helpless at birth. The mother stays with them and defends them fiercely from any animal that may approach—even their own father. Young jaguars learn to hunt by living with their mothers for two years or more.

Jaguars are still hunted for their attractive fur. Ranchers also kill them because the cats sometimes prey upon their livestock.

The Jaguar is the largest cat in the Western Hemisphere and the third largest cat in the world (after the Lion and the Tiger.) It is also one of the four roaring cats. It differs from a lion's roar and is more of a series of hoarse coughs. It is often confused with the leopard but the Jaguar is a stockier animal. It is usually larger with a broad head and shorter legs and tail. The color is generally a tawny yellow with dark spots on the head and neck and dark rings on the body. Inside these rings there is usually a dark spot. This is the primary difference between the spots on a Jaguar and the spots on a leopard. There are also black Jaguars. These are usually found in dense forests and are often called Black Panthers. The body length is between 4 and 6 feet and its tail is about 30 inches.

5. Quais tags você atribuiria a este documento?

The origins of Jaguar can be traced back to the northern seaside town of Blackpool in the early 1920s. It was here that a young motorcycle enthusiast, Bill Lyons (b. 1901), not yet 21 years of age, met William Walmsley (b. 1891) who was building attractive motorcycle sidecars and attaching them to reconditioned motorbikes. Walmsley had not long arrived in Blackpool with his parents from Stockport, and both families happened lived in the same street – King Edward Avenue.

As soon as William Lyons came of age, he and Walmsley formed the Swallow Sidecar Company on 4th September 1922 with a bank overdraft of £1,000. Securing first and second floor premises in Bloomfield Road, Blackpool, they commenced commercial production of the sidecars together with a small team of eight employees, including a young Arthur Whitaker. Although initially employed to help with sales, Whitaker's strength lay in purchasing and he was to remain with Lyons for some 50 years, proving himself to be one of the most shrewd purchasers in the business.

Lack of factory space soon became a problem, and two further Blackpool sites were taken over – in Woodfield Road (mainly for despatch purposes) and, shortly afterwards, in John Street which was fortuitously situated close to the main Swallow premises.

In mid-1926, plans for producing motor-car bodies were well under way, and this – together with the year-by-year increase in production of the sidecars – made it necessary for Swallow to move into a larger building. Lyons had heard that a building erected specifically for coachbuilding was coming on to the market. The previous occupant, Joseph Street, had run into trouble and the property was now up for sale, but at a price beyond which the partnership could afford. Fortunately, Walmsley's father had just sold his coal business and was looking for somewhere to invest the proceeds,

offering to purchase the building and lease it to Lyons and Walmsley junior at an annual rent of £325.

The entire removal to 41 Cocker Street took just one weekend with no assistance from outside sources, other than the unofficial assistance of one pantechnicon and driver, which on the Friday had delivered new sidecar chassis frames from Haywards of Birmingham.

It was in late 1926, and announced to the public in May 1927, that the Swallow Sidecar and Coachbuilding Company first diversified by taking an existing car and bodying it with more fashionable coachwork. The first model to benefit was the popular, but basic, Austin Seven. Intended to bring motoring to the masses, the Austins were cheap and easy to run, but Lyons believed "… that it would also appeal to a lot of people if it had a more luxurious and attractive body." (…)

6. Quais tags você atribuiria a este documento?

Banco de dados, é um conjunto de registros dispostos em estrutura regular que possibilita a reorganização dos mesmos e produção de informação. Um banco de dados normalmente agrupa registros utilizáveis para um mesmo fim.

Um banco de dados é usualmente mantido e acessado por meio de um software conhecido como Sistema Gerenciador de Banco de Dados (SGBD). Normalmente um SGBD adota um modelo de dados, de forma pura, reduzida ou estendida. Muitas vezes o termo banco de dados é usado, de forma errônea, como sinônimo de SGDB. O modelo de dados mais adotado hoje em dia é o modelo relacional, onde as estruturas têm a forma de tabelas, compostas por tuplas (linhas) e colunas.

Os bancos de dados são utilizados em muitas aplicações, abrangendo praticamente todo o campo dos programas de computador. Os bancos de dados são o método de armazenamento preferencial e baseiam-se em tecnologias padronizadas de bancos de dados.

Um banco de dados é um conjunto de informações com uma estrutura regular. Um banco de dados é normalmente, mas não necessariamente, armazenado em algum formato de máquina legível para um computador. Há uma grande variedade de bancos de dados, desde simples tabelas armazenadas em um único arquivo até gigantescos bancos de dados com muitos milhões de registos, armazenados em salas cheias de discos rígidos.

A apresentação dos dados geralmente é semelhante à de uma planilha eletrônica, porém os sistemas de gestão de banco de dados possuem características especiais para o armazenamento, classificação, gestão da integridade e recuperação dos dados. Com a evolução de padrões de conectividade entre as tabelas de um banco de dados e programas desenvolvidos em linguagens como Java, Delphi, Visual Basic, C++ etc, a

apresentação dos dados, bem como a navegação, passou a ser definida pelo programador ou o designer de aplicações. Como hoje em dia a maioria das linguagens de programação fazem ligações a bancos de dados, a apresentação destes tem ficado cada vez mais a critério dos meios de programação, fazendo com que os bancos de dados deixem de restringir-se às pesquisas básicas, dando lugar ao compartilhamento, em tempo real, de informações, mecanismos de busca inteligentes e permissividade de acesso hierarquizada.

7. Quais tags você atribuiria a este documento?

Computational intelligence is the study of the design of intelligent agents. An agent is something that acts in an environment—it does something. Agents include worms, dogs, thermostats, airplanes, humans, organizations, and society. An intelligent agent is a system that acts intelligently: What it does is appropriate for its circumstances and its goal, it is flexible to changing environments and changing goals, it learns from experience, and it makes appropriate choices given perceptual limitations and finite computation.

The central scientific goal of computational intelligence is to understand the principles that make intelligent behavior possible, in natural or artificial systems. The main hypothesis is that reasoning is computation. The central engineering goal is to specify methods for the design of useful, intelligent artifacts.

Artificial intelligence (AI) is the established name for the field we have defined as computational intelligence (CI), but the term “artificial intelligence” is a source of much confusion. Is artificial intelligence real intelligence? Perhaps not, just as an artificial pearl is a fake pearl, not a real pearl. “Synthetic intelligence” might be a better name, since, after all, a synthetic pearl may not be a natural pearl but it is a real pearl. However, since we claimed that the central scientific goal is to understand both natural and artificial (or synthetic) systems, we prefer the name “computational intelligence.” It also has the advantage of making the computational hypothesis explicit in the name.

The confusion about the field’s name can, in part, be attributed to a confounding of the field’s purpose with its methodology. The purpose is to understand how intelligent behavior is possible. The methodology is to design, build, and experiment with computational systems that perform tasks commonly viewed as intelligent.

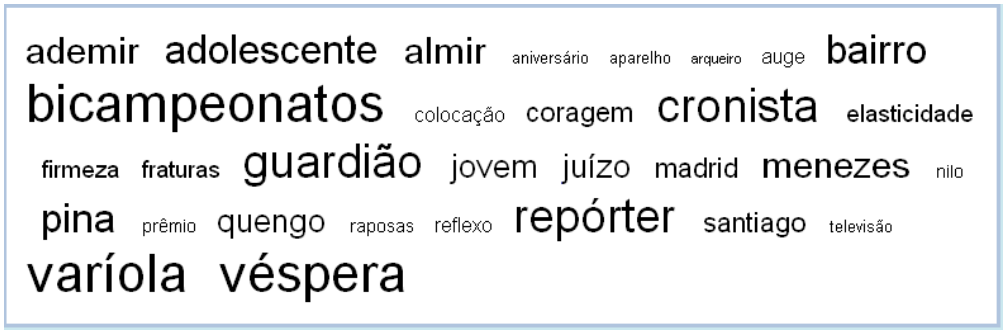
Building these artifacts is an essential activity since computational intelligence is, after all, an empirical science; but it shouldn't be confused with the scientific purpose.

Another reason for eschewing the adjective "artificial" is that it connotes simulated intelligence. Contrary to another common misunderstanding, the goal is not to simulate intelligence. The goal is to understand real (natural or synthetic) intelligent systems by synthesizing them. A simulation of an earthquake isn't an earthquake; however, we want to actually create intelligence, as you could imagine creating an earthquake.

Questões 8 a 10: Em cada questão é apresentada uma *Tag Cloud Diferencial* e pede-se para relacioná-la ao documento que a originou. O documento será algum dos 7 documentos que foram apresentados nas primeiras 7 questões.

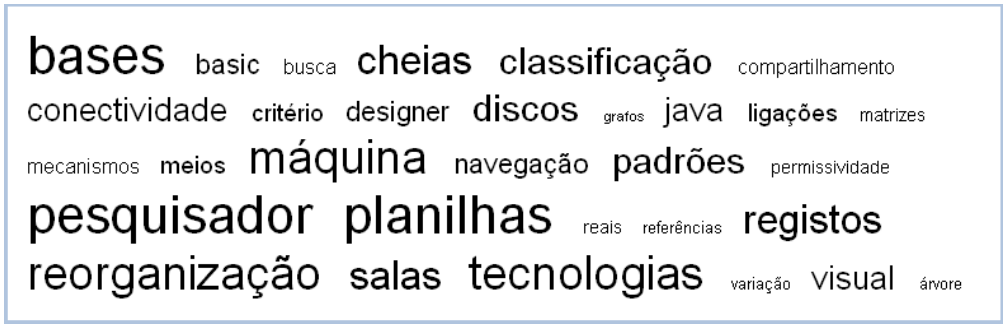
---

8. Relacione a Tag Cloud Diferencial com um dos textos apresentados na questão de 1 a 7. Identifique o texto pelo número da questão.



Esta ***Tag Cloud Diferencial*** está relacionada com o texto da questão: ( )

9. Relacione a Tag Cloud Diferencial com um dos textos apresentados na questão de 1 a 7. Identifique o texto pelo número da questão.



Esta ***Tag Cloud Diferencial*** está relacionada com o texto da questão: ( )

10. Relacione a Tag Cloud Diferencial com um dos textos apresentados na questão de 1 a 7. Identifique o texto pelo número da questão.

Esta ***Tag Cloud Diferencial*** está relacionada com o texto da questão: ( )

Questões 11 a 16: Em cada questão é apresentada uma tag cloud e você deve avaliar a representatividade da tag cloud em relação ao assunto do documento ao qual ela se refere.

Você pode escolher uma das avaliações possíveis: {Concordo totalmente, Concordo parcialmente, Nem concordo nem discordo, Discordo parcialmente, Discordo totalmente}

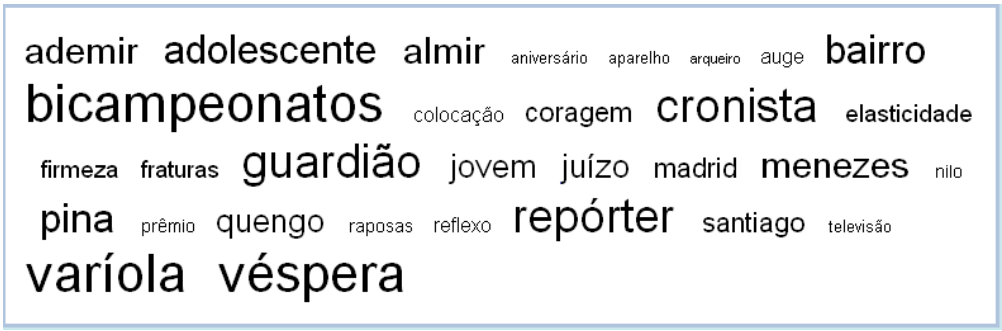
Por representatividade entende-se que as *tags* apresentadas na *tag cloud* tenham alguma relação com o assunto citado na questão.

---

11. Avalie a seguinte tag cloud gerada para um documento sobre o assunto Manga (Fruta).

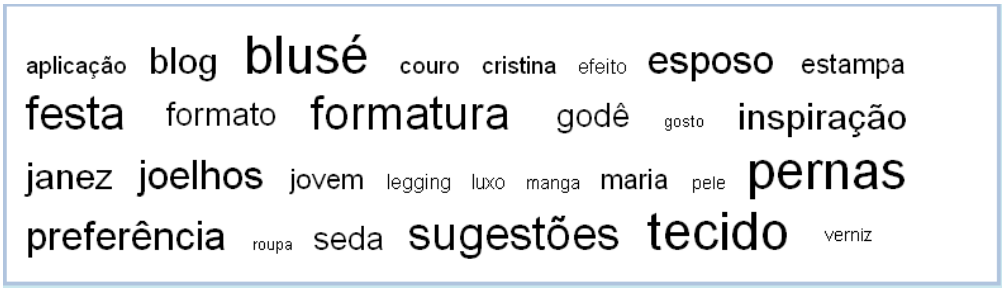
( ) Concordo totalmente
( ) Concordo parcialmente
( ) Nem concordo nem discordo
( ) Discordo parcialmente
( ) Discordo totalmente

12. Avalie a seguinte tag cloud gerada para um documento sobre o assunto Manga (Jogador).



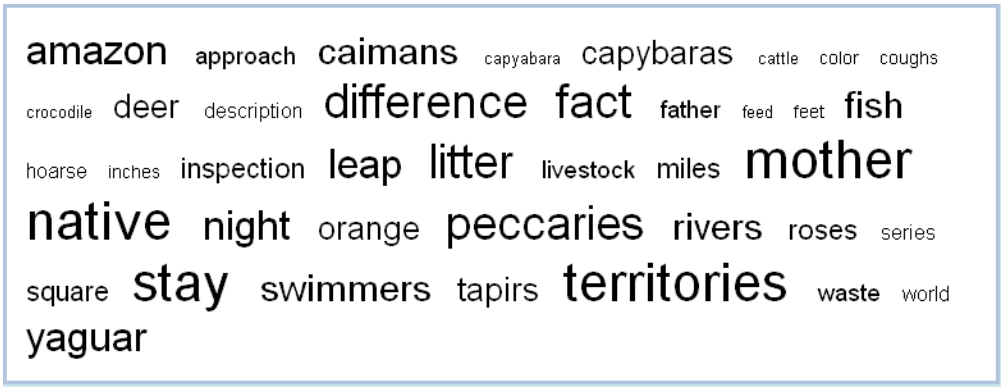
( ) Concordo totalmente
( ) Concordo parcialmente
( ) Nem concordo nem discordo
( ) Discordo parcialmente
( ) Discordo totalmente

13. Avalie a seguinte tag cloud gerada para um documento sobre o assunto Manga (Roupa).



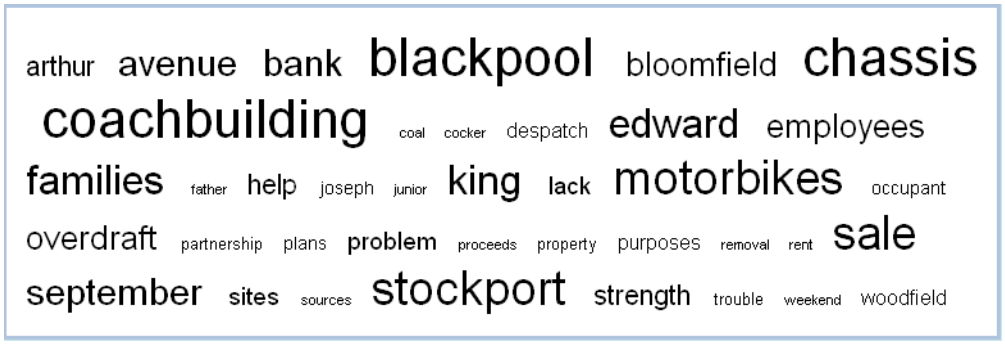
( ) Concordo totalmente
( ) Concordo parcialmente
( ) Nem concordo nem discordo
( ) Discordo parcialmente
( ) Discordo totalmente

14. Avalie a seguinte tag cloud gerada para um documento sobre o assunto Jaguar (Animal).



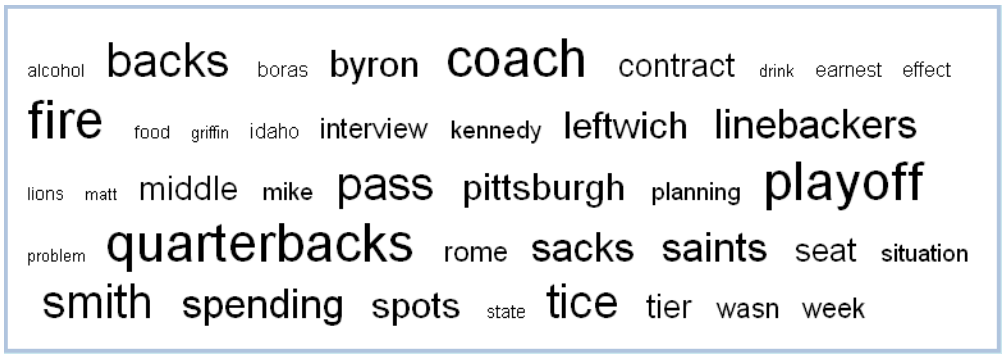
( ) Concordo totalmente
( ) Concordo parcialmente
( ) Nem concordo nem discordo
( ) Discordo parcialmente
( ) Discordo totalmente

15. Avalie a seguinte tag cloud gerada para um documento sobre o assunto Jaguar (Carro).



( ) Concordo totalmente
( ) Concordo parcialmente
( ) Nem concordo nem discordo
( ) Discordo parcialmente
( ) Discordo totalmente

16. Avalie a seguinte tag cloud gerada para um documento sobre o assunto Jaguar (Esporte).



( ) Concordo totalmente
( ) Concordo parcialmente
( ) Nem concordo nem discordo
( ) Discordo parcialmente
( ) Discordo totalmente

Questões 17 e 18: Em cada questão pede-se para avaliar a representatividade de uma *Tag Cloud de Resumo do Conjunto* sobre o assunto específico de cada questão. Você pode escolher uma das avaliações possíveis: {Concordo totalmente, Concordo parcialmente, Nem concordo nem discordo, Discordo parcialmente, Discordo totalmente}

Por representatividade entende-se que as *tags* apresentadas na *tag cloud* tenham alguma relação com o assunto citado na questão.

---

17. O assunto em questão é Manga (Fruta). Qual avaliação você atribui a seguinte tag cloud de resumo do conjunto.

( ) Concordo totalmente

( ) Concordo parcialmente

( ) Nem concordo nem discordo

( ) Discordo parcialmente

( ) Discordo totalmente

18. O assunto em questão é Jaguar (Jogo). Qual avaliação você atribui a seguinte tag cloud de resumo do conjunto.

( ) Concordo totalmente

( ) Concordo parcialmente

( ) Nem concordo nem discordo

( ) Discordo parcialmente

( ) Discordo totalmente