TAGS

uso de texto não-hierárquico para a classificação de conteúdo

MAC 5855

Nelson Lago
Márcio Hasegawa

Organização de conteúdo

- **Organização hierárquica**
  - Similar à organização de diretórios no sistema de arquivos
  - Navegação recursiva
  - Relação entre os itens é hierárquica
  - Categorização é necessariamente manual e trabalhosa
  - Cada item só pode ser incluído em uma categoria, mas nem sempre é claro qual é a categoria adequada

Organização de conteúdo

- **Organização com etiquetas (*tags*)**
  - Um único nível
  - Navegação baseada em similaridade
    - o sistema pode criar estruturas de navegação automaticamente em função do contexto
  - Relação entre os itens é baseada no relacionamento entre conceitos
  - Categorização pode ser automatizada
  - Um item pode pertencer a diversas categorias

Definições

- ***Tags* podem ser palavras isoladas ou frases curtas**
- ***Tags* podem ser associadas a qualquer tipo de conteúdo**
  - Usuários
  - Produtos
  - Textos, fotos ou vídeos
  - Favoritos
- ***Tags* devem ser tratadas pelas raízes das palavras**
  - Plurais devem ser eliminados
  - A caixa deve ser igualada
  - Se possível, sinônimos importantes devem ser identificados

Definições

- **O vocabulário de uma aplicação é o conjunto de todas as *tags* registradas**
  - Se o mesmo vocabulário é usado para descrever usuários e itens de interesse dos usuários, é possível identificar os itens de maior relevância para cada usuário
  - Da mesma maneira, é possível identificar usuários com perfis de interesse similares, bem como itens de apelos similares
- **A partir do comportamento do usuário, é possível:**
  - identificar conteúdos relacionados
  - identificar *tags* relacionadas entre si
  - identificar grupos de usuários com interesses comuns

Usos

- **Usuários criam *tags* para organizar dados de seu interesse**
  - Organização individualizada, e não dependente da organização preferida pelo autor do conteúdo
  - Facilita a interação entre usuários
    - Identificar usuários com interesses semelhantes
    - Identificar conteúdos de interesse de acordo com outros usuários
  - Navegação baseada em conceitos
- **A partir das tags, é possível extrair diversas informações, em particular relações de semelhança que podem ser usadas para recomendações, navegação, ordenação de relevância etc.**

Vetores de termos

- ***Tags* são um mecanismo para a criação de metadados sobre um dado qualquer na forma de um vetor de termos**
  - Um vetor de termos consiste em uma lista de pesos, ou “medidas de relevância”, de todos os termos do vocabulário para um determinado item
  - Pode ser visto também como uma matriz esparsa de duas dimensões, com todos os itens nas linhas e todos os termos nas colunas; cada elemento da matriz é a relevância do termo para o item
  - Mantendo-se os dados normalizados, é possível computar a distância entre dois vetores de termos e, portanto, entre dois itens

Criação de *tags*

- **Três maneiras de criar *tags*:**
  - Geração profissional
  - Geração colaborativa pelos usuários
  - Geração automatizada

Criação de *tags*: geração profissional

- ***Tags* geradas manualmente por profissionais:**
  - Em geral, são altamente relevantes para os conteúdos correspondentes
  - Independem das palavras usadas explicitamente nesses conteúdos
  - Representam conceitos relevantes para o domínio da aplicação, em função do conhecimento profissional
  - Podem constituir um vocabulário controlado

Criação de *tags*: geração profissional

- **Por outro lado:**
  - Envolvem alto custo de criação
  - Só são de fato viáveis com conteúdos que não são criados pelos usuários
  - Só fazem sentido em ambientes onde o escopo da aplicação é restrito, como sites de interesse em tópicos específicos
  - Podem utilizar vocabulário incompatível com os usuários

Criação de *tags*: geração colaborativa

- ***Tags* geradas colaborativamente pelos usuários:**
  - Têm custo praticamente nulo
  - Permitem ao usuário criar suas próprias *tags*
  - Independem das palavras usadas explicitamente nos conteúdos
  - Geralmente, representam conceitos relevantes para o domínio da aplicação
  - Permitem a identificação de dados adicionais sobre os usuários e os conteúdos, de acordo com o comportamento dos usuários

Criação de *tags*: geração colaborativa

- **Por outro lado:**
  - Precisam ser filtradas para a identificação de *tags* iguais mas escritas de forma diferente ou de sinônimos
  - Precisam ser filtradas para a eliminação de “lixo”
  - Precisam de um volume relativamente alto de usuários aplicando *tags* aos conteúdos para que algumas formas de uso sejam viabilizadas

Criação de *tags*: geração automatizada

- ***Tags* geradas automaticamente:**
  - Têm baixo custo
  - Podem lidar com grandes volumes de dados que mudam constantemente
  - Independem da participação dos usuários
  - Podem ser aplicadas a conteúdos de escopo indeterminado

Criação de *tags*: geração automatizada

- **Por outro lado:**
  - São restritas às palavras contidas no próprio texto
    - exceto pela inserção manual de alguns sinônimos
  - Nem sempre representam conceitos relevantes para o domínio da aplicação
  - Geralmente são restritas a uma palavra por *tag*
  - Podem gerar muito "lixo"
    - palavras com sentido dependente de contexto, como "ganho"
    - palavras de pouca relevância

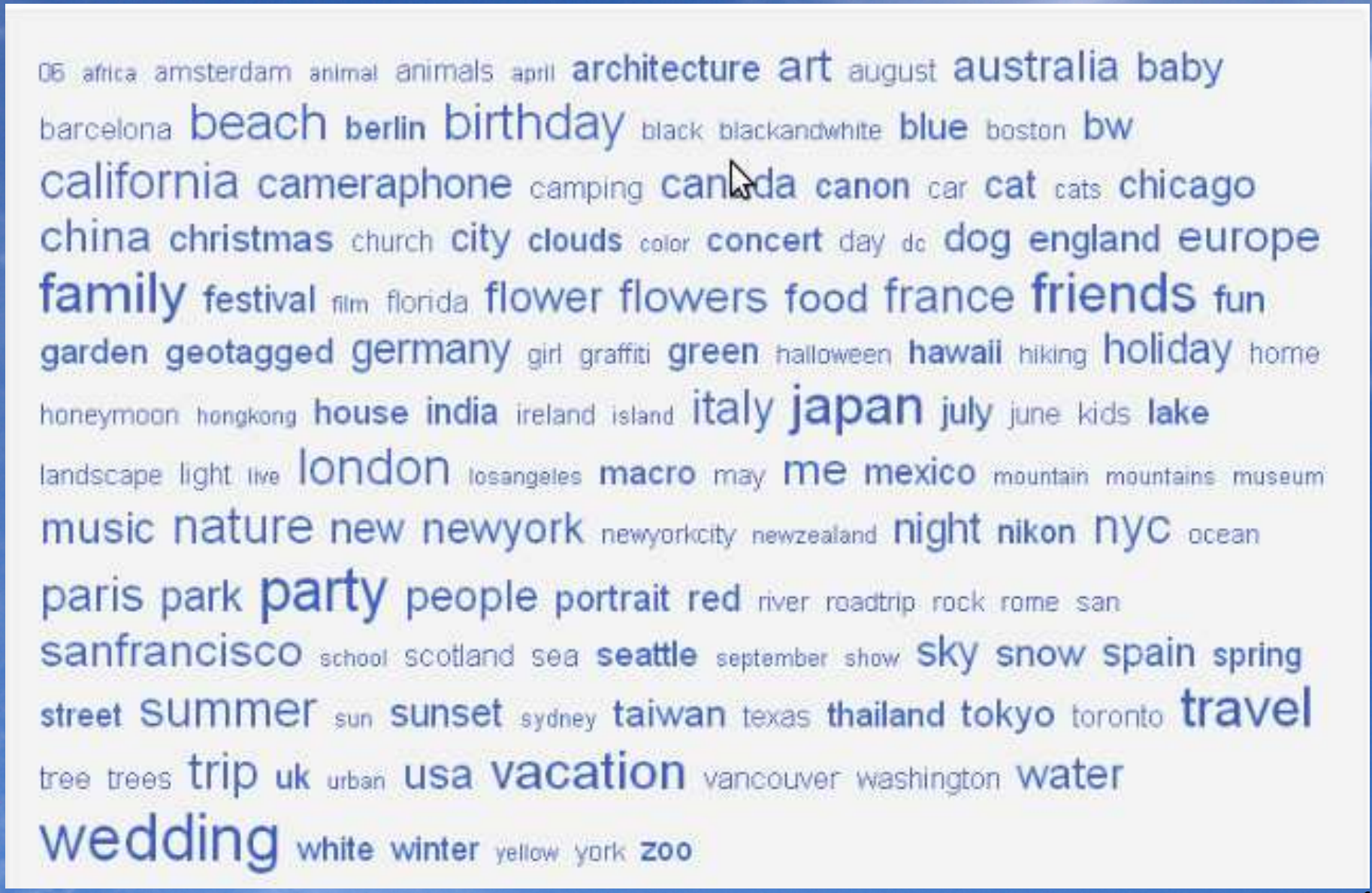
Criação de *tags*: sugestões

- **Algumas sugestões para maximizar a eficiência de um sistema de *tags*:**
  - Criar dicionários
    - um dicionário de sinônimos permite identificar tags iguais ou conteúdos relacionados
    - um dicionário de frases permite contextualizar conceitos (como “ganho de capital”) e aumentar a relevância das *tags*
    - A análise da co-ocorrência de tags pode auxiliar na identificação automática de tags que são sinônimos
  - Para tags geradas automaticamente, utilize poucas *tags* altamente relevantes
  - Processe as palavras para basear as *tags* em suas raízes
  - As três formas de geração de *tags* podem ser combinadas

Navegação: nuvens de *tags*

- ***Tags*** **podem ser usadas para navegação**
- **A forma de apresentação mais comum nesse caso é a nuvem de *tags***
- **O autor apresenta uma implementação completa de uma nuvem de *tags* no livro**

Navegação: nuvens de *tags*



Implementação: Correlação entre itens

- **Calculando a correlação entre 2 itens utilizando tags**
  1. Verificar quais as tags foram aplicadas aos itens
  2. Verificar quantas vezes as tags foram aplicadas a cada item
  3. Normalizar as informações
  4. Calcular a correlação

Implementação: Exemplo

	Tag 1	Tag 2	Tag 3	Tag 4	Tag 5	Normalizador
Item 1	4	8	6	3	0	11,18
Item 2	0	5	0	8	5	10,68

$\text{Math.sqrt}(4^2 + 8^2 + 6^2 + 3^2) = 11,18$

	Tag 1	Tag 2	Tag 3	Tag 4	Tag 5
Item 1	0,3578	0,7156	0,5367	0,2683	0
Item 2	0	0,4682	0	0,7491	0,4682

4 / 11,18 = 0,3578

Correlação = (0,7156 * 0,4682) +
(0,2683 * 0,7491)
= 0,5360

Exemplo: MyRank

- **O Yahoo! usa *tags* para identificar a relevância de buscas**
  - Usuários podem aplicar *tags* a páginas de seu interesse
  - Um alto número de *tags* similares aplicadas a uma mesma página sugere uma maior relevância da página para o termo
  - Os resultados da busca são individualizados e influenciados pelas *tags* que o usuário e seus "vizinhos" mais usam
- **O mecanismo é diferente, mas o princípio é o mesmo do PageRank**

Exemplo: Music Explaura

- **Projeto de pesquisa da Sun sobre recomendações automáticas**
  - Sistema de recomendações musicais baseado em *tags* extraídas automaticamente através de *data mining* na *web*
  - *Tags* textuais são usadas para orientar as recomendações, e os usuários podem ver essas *tags* em uma nuvem de *tags*
  - Usuário pode interagir com a relevância das *tags* (diretamente na nuvem de *tags*) para modificar os resultados das recomendações

Exemplo: Music Explaura

Caetano Veloso

portuguese latin 60s political rock brasileiro folk jazz world music 70s bra

mpb brazil psychedelic samba tropicalia

brazilian world rock

By Luiza Leite.

Caetano Emanuel Viana Telles Veloso (born August 7, 19
Caetano Veloso, is a composer, singer, guitarist, writer, a
He has been called "one of the greatest songwriters of th
sometimes considered to be the Bob Dylan of Brazil. Velo
his participation in the Brazilian musical movement Tropi
encompassed theatre, poetry and music in the 1960s, at t
Brazilian military dictatorship. Veloso was born in Bahia,
northeastern area of Brazil, but moved to Rio de Janeiro
the mid-1960s. Soon after the move, Veloso won a music
to his first label. He became one of the founders of Trop
of several other musicians and artists—including his sister
same period. However the Brazilian government at the ti
music and political action as threatening, and he was arr

Similar artists

All

Chico Buarque why? diff
Tags: mpb, samba, brasil, brazilian
Popularity:

Elis Regina why? diff
Tags: mpb, brazilian, brasil, brazil
Popularity:

Jorge Ben why? diff
Tags: mpb, bossa nova, brazilian, samba
Popularity:

Gal Costa why? diff
Tags: tropicalia, mpb, brazilian, brasil
Popularity:

Maria Bethânia why? diff
Tags: mpb, bossa nova, brasil, brazilian
Popularity: