Classificando Automaticamente Documentos Digitais no *Site* de Notícias do UOL

Elias Oliveira, Patrick Marques Ciarelli,

Marcos Hercules Santos e Bruno Oliveira da Costa

elias@inf.ufes.br

Laboratório de Recuperação *Inteligente* da Informação

Departamento de Ciências da Informação

Universidade Federal do Espírito Santo

Campus de Goiabeiras, Av. Fernando Ferrari, s/n, Cx Postal 5011 29060-970

Projeto parcialmente financiado pelo FACITEC

Roteiro da Apresentação

- Motivação;
- Algumas Práticas;
- Nossa Proposta;
- Experimentos;
- Conclusões.

Motivação (I)

- Crescente número de publicações.
  O GOOGLE indexa **8 bilhões** de páginas;

Motivação (I)

- Crescente número de publicações.
  O GOOGLE indexa **8 bilhões** de páginas;
- Grande número de áreas, e sub-áreas, do conhecimento;

Motivação (I)

- Crescente número de publicações.
  O GOOGLE indexa **8 bilhões** de páginas;
- Grande número de áreas, e sub-áreas, do conhecimento;
- Dinâmica perceptiva da sociedade;

Motivação (I)

- Crescente número de publicações.
  O GOOGLE indexa **8 bilhões** de páginas;
- Grande número de áreas, e sub-áreas, do conhecimento;
- Dinâmica perceptiva da sociedade;
- A demanda cresce cada vez mais por **porções de informação**;

Motivação (I)

- Crescente número de publicações.
  O GOOGLE indexa **8 bilhões** de páginas;
- Grande número de áreas, e sub-áreas, do conhecimento;
- Dinâmica perceptiva da sociedade;
- A demanda cresce cada vez mais por **porções de informação**;
- ...

Motivação (II)

Diante dessa multitude de documentos espalhados na rede Internet e, mais recentemente, em Bibliotecas Digitais...

- Como recuperar, com boa precisão, as **porções de informação** de interesse de um particular usuário?

Motivação (II)

Diante dessa multitude de documentos espalhados na rede Internet e, mais recentemente, em Bibliotecas Digitais...

- Como recuperar, com boa precisão, as **porções de informação** de interesse de um particular usuário?
- Como detectar um novo nicho de interesse de informação;

Motivação (II)

Diante dessa multitude de documentos espalhados na rede Internet e, mais recentemente, em Bibliotecas Digitais...

- Como recuperar, com boa precisão, as **porções de informação** de interesse de um particular usuário?
- Como detectar um novo nicho de interesse de informação;
- Quem escreveu algo parecido, ou no mesmo assunto?

Motivação (II)

Diante dessa multitude de documentos espalhados na rede Internet e, mais recentemente, em Bibliotecas Digitais...

- Como recuperar, com boa precisão, as **porções de informação** de interesse de um particular usuário?
- Como detectar um novo nicho de interesse de informação;
- Quem escreveu algo parecido, ou no mesmo assunto?
- Catálogos mais **inteligentes**, **dinâmicos** e **auto-configuráveis segundo os Interesses do Usuário**;

Motivação (II)

Diante dessa multitude de documentos espalhados na rede Internet e, mais recentemente, em Bibliotecas Digitais...

- Como recuperar, com boa precisão, as **porções de informação** de interesse de um particular usuário?
- Como detectar um novo nicho de interesse de informação;
- Quem escreveu algo parecido, ou no mesmo assunto?
- Catálogos mais **inteligentes**, **dinâmicos** e **auto-configuráveis segundo os Interesses do Usuário**;
- ...

Algumas Práticas...

A efetiva **leitura**, **análise** e a **interpretação** do conteúdo dos documentos tornou-se um processo extremamente caro.

- Então o que precisamos fazer? – Precisamos de **uma nova metodologia** para:

Algumas Práticas...

A efetiva **leitura**, **análise** e a **interpretação** do conteúdo dos documentos tornou-se um processo extremamente caro.

- Então o que precisamos fazer? – Precisamos de **uma nova metodologia** para:
  1. Representação;

Algumas Práticas...

A efetiva **leitura**, **análise** e a **interpretação** do conteúdo dos documentos tornou-se um processo extremamente caro.

- Então o que precisamos fazer? – Precisamos de **uma nova metodologia** para:
  1. Representação;
  2. Indexação e, posteriormente,

Algumas Práticas...

A efetiva **leitura**, **análise** e a **interpretação** do conteúdo dos documentos tornou-se um processo extremamente caro.

- Então o que precisamos fazer? – Precisamos de **uma nova metodologia** para:
  1. Representação;
  2. Indexação e, posteriormente,
  3. Classificação das **porções de informação**.

Algumas Práticas...

A efetiva **leitura**, **análise** e a **interpretação** do conteúdo dos documentos tornou-se um processo extremamente caro.

- Então o que precisamos fazer? – Precisamos de **uma nova metodologia** para:
  1. Representação;
  2. Indexação e, posteriormente,
  3. Classificação das **porções de informação**.
- Precisamos que a máquina nos auxilie mais do que simplesmente armazenar registros;

Algumas Práticas...

A efetiva **leitura**, **análise** e a **interpretação** do conteúdo dos documentos tornou-se um processo extremamente caro.

- Então o que precisamos fazer? – Precisamos de **uma nova metodologia** para:
  1. Representação;
  2. Indexação e, posteriormente,
  3. Classificação das **porções de informação**.
- Precisamos que a máquina nos auxilie mais do que simplesmente armazenar registros;
- Precisamos de um assistente automático que nos ajude a encontrar, analisar e agrupar **porções de informação**, ou documentos, semelhantes sob os mais diversos aspectos.

Algumas Práticas...

A efetiva **leitura**, **análise** e a **interpretação** do conteúdo dos documentos tornou-se um processo extremamente caro.

- Então o que precisamos fazer? – Precisamos de **uma nova metodologia** para:
  1. Representação;
  2. Indexação e, posteriormente,
  3. Classificação das **porções de informação**.
- Precisamos que a máquina nos auxilie mais do que simplesmente armazenar registros;
- Precisamos de um assistente automático que nos ajude a encontrar, analisar e agrupar **porções de informação**, ou documentos, semelhantes sob os mais diversos aspectos.

Representação Vetorial do Documento

Nós seres humanos "pensamos", as máquinas "fazem contas"...

Representação Vetorial do Documento

Nós seres humanos "pensamos", as máquinas "fazem contas"...

Precisamos transformar o processo de classificação em um processo de contagem...

Representação Vetorial do Documento

Nós seres humanos "pensamos", as máquinas "fazem contas"...
Precisamos transformar o processo de classificação em um processo de contagem...

Vamos supor que tenhamos uma base de dados $D = \{d_1, d_2, \ldots, d_j, \ldots, d_n\}$ e queiramos saber quão similar q (um outro documento) é de um ou mais documentos em D.

Representação Vetorial do Documento

Nós seres humanos "pensamos", as máquinas "fazem contas"...
Precisamos transformar o processo de classificação em um processo de contagem...

Vamos supor que tenhamos uma base de dados $D = \{d_1, d_2, \ldots, d_j, \ldots, d_n\}$ e queiramos saber quão similar q (um outro documento) é de um ou mais documentos em D.

$$sim(d_j, q) = \frac{\mathbf{d_j} \bullet \mathbf{q}}{|\ \mathbf{d_j}\ | \times |\ \mathbf{q}\ |}$$

Representação Vetorial do Documento

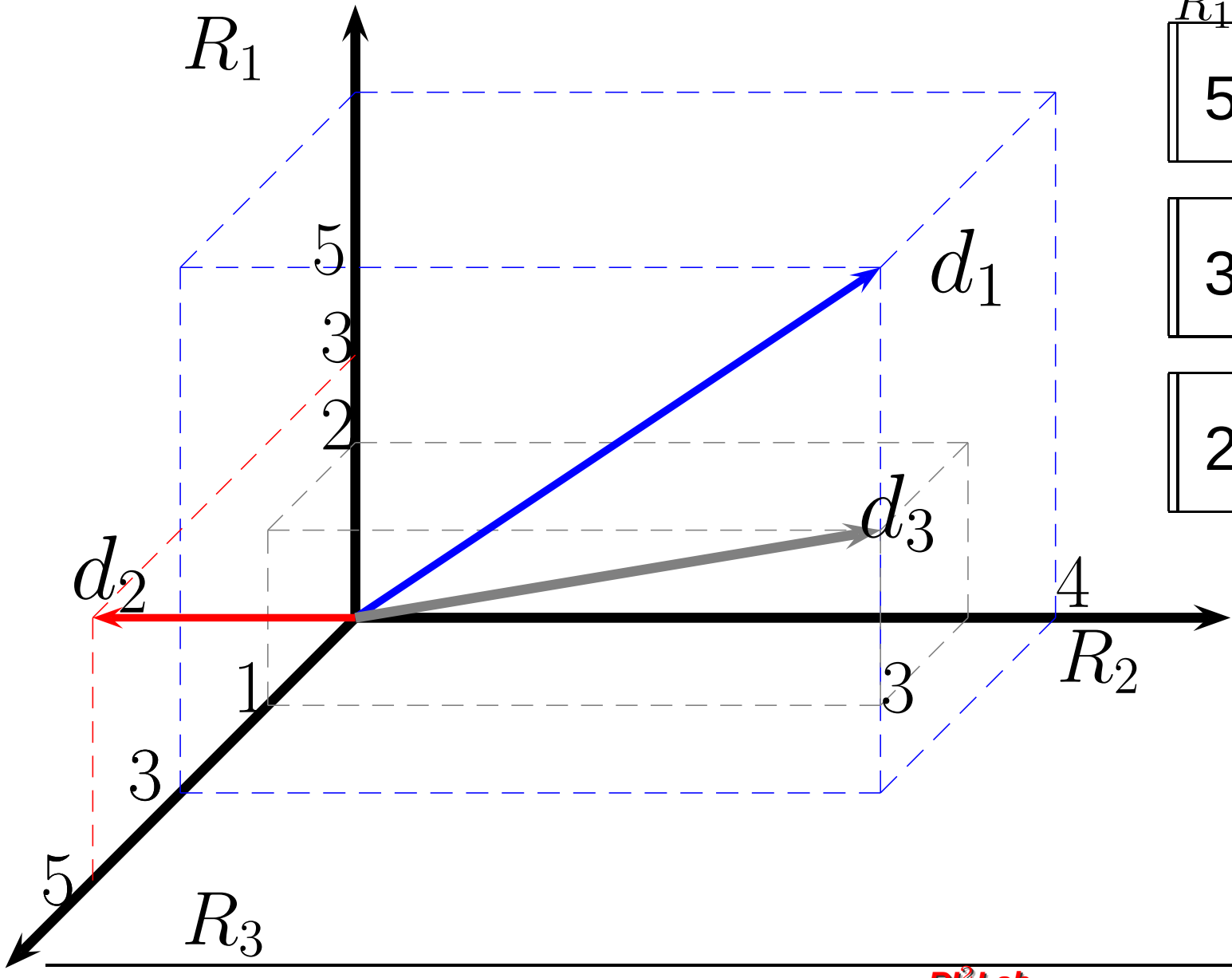
Nós seres humanos "pensamos", as máquinas "fazem contas"...
Precisamos transformar o processo de classificação em um processo de contagem...

Vamos supor que tenhamos uma base de dados $D = \{d_1, d_2, \ldots, d_j, \ldots, d_n\}$ e queiramos saber quão similar q (um outro documento) é de um ou mais documentos em D.

$$sim(d_j, q) = \frac{\mathbf{d_j} \bullet \mathbf{q}}{|\ \mathbf{d_j}\ | \times |\ \mathbf{q}\ |}$$

$$= \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{j=1}^{t} w_{i,q}^2}} = cos(\theta)$$

Visualização de Documentos



R_1	R_2	R_3	
5	4	3	d_1
3	0	5	d_2
2	3	1	d_3

Ilustração da Metodologia (I)

Para validação da metodologia proposta, utilizamos alguns *documentos* de notícias do site UOL.

Ilustração da Metodologia (I)

Para validação da metodologia proposta, utilizamos alguns *documentos* de notícias do site UOL.

Esses documentos já são classificados pelos especialistas humano. Isso nos ajudará a validarmos nossa metodologia.

Ilustração da Metodologia (I)

Para validação da metodologia proposta, utilizamos alguns *documentos* de notícias do site UOL.

Esses documentos já são classificados pelos especialistas humano. Isso nos ajudará a validarmos nossa metodologia.

Para que o **modelo** possa **aprender** é preciso que o especialista humano apresente uma boa quantidade de **documentos corretamente classificados**.

Ilustração da Metodologia (II)

	cinema	economia	esporte
cin1	0.499	0.0	0.0
cin2	0.417	0.0	0.039
cin3	0.408	0.0	0.0
cin4	0.512	0.0	0.0
cin5	0.399	0.053	0.0
eco1	0.0	0.415	0.0
eco2	0.0	0.626	0.0
eco3	0.0	0.357	0.0
eco4	0.0	0.409	0.0
eco5	0.0	0.643	0.0
esp1	0.0	0.0	0.419
esp2	0.0	0.0	0.418
esp3	0.03	0.0	0.467
esp4	0.0	0.01	0.552
esp5	0.0	0.0	0.370

Ilustração da Metodologia (III)

Ponto de Corte para as Classes

	cinema	economia	esporte
pc:	0.399	0.370	0.384

Cálculo dos valores de ponto de corte para cada uma das classes consideradas nos experimentos.

Ilustração da Metodologia (IV)

	Classes de Documentos		
	cinema	economia	esporte
cin6	0.296	0.000	0.000
eco6	0.000	0.575	0.000
esp6	0.120	0.000	0.541
pc:	**0.399**	**0.370**	**0.384**

Cálculo de similaridade entre os documentos analisados e os discriminantes das classes.

Conclusões

- Crescimento vertiginoso de repositórios de informação no Brasil;

Conclusões

- Crescimento vertiginoso de repositórios de informação no Brasil;
- Os métodos tradicionais de tratamento da informação não são mais compatíveis com repositórios do tamanho da Internet;

Conclusões

- Crescimento vertiginoso de repositórios de informação no Brasil;
- Os métodos tradicionais de tratamento da informação não são mais compatíveis com repositórios do tamanho da Internet;
- Portanto, **devemos buscar novas soluções**.

Conclusões

- Crescimento vertiginoso de repositórios de informação no Brasil;
- Os métodos tradicionais de tratamento da informação não são mais compatíveis com repositórios do tamanho da Internet;
- Portanto, **devemos buscar novas soluções**.
- Os documentos sendo representados através de vetores, nos permite utilizar o cálculo do ângulo entre vetores como medida de similaridade entre quaisquer dois documentos;

Conclusões

- Crescimento vertiginoso de repositórios de informação no Brasil;
- Os métodos tradicionais de tratamento da informação não são mais compatíveis com repositórios do tamanho da Internet;
- Portanto, **devemos buscar novas soluções**.
- Os documentos sendo representados através de vetores, nos permite utilizar o cálculo do ângulo entre vetores como medida de similaridade entre quaisquer dois documentos;
- Testamos o modelo apresentado em um conjunto de documentos já previamente classificado pelo especialista humano;

Conclusões

- Crescimento vertiginoso de repositórios de informação no Brasil;
- Os métodos tradicionais de tratamento da informação não são mais compatíveis com repositórios do tamanho da Internet;
- Portanto, **devemos buscar novas soluções**.
- Os documentos sendo representados através de vetores, nos permite utilizar o cálculo do ângulo entre vetores como medida de similaridade entre quaisquer dois documentos;
- Testamos o modelo apresentado em um conjunto de documentos já previamente classificado pelo especialista humano;
- Os resultados nos mostraram que o modelo trouxe, **de forma automática**, a mesma classificação dada pelo especialista humano;

Conclusões

- Crescimento vertiginoso de repositórios de informação no Brasil;
- Os métodos tradicionais de tratamento da informação não são mais compatíveis com repositórios do tamanho da Internet;
- Portanto, **devemos buscar novas soluções**.
- Os documentos sendo representados através de vetores, nos permite utilizar o cálculo do ângulo entre vetores como medida de similaridade entre quaisquer dois documentos;
- Testamos o modelo apresentado em um conjunto de documentos já previamente classificado pelo especialista humano;
- Os resultados nos mostraram que o modelo trouxe, **de forma automática**, a mesma classificação dada pelo especialista humano;
- E agora?