Indexação, classificação e buscas

Matheus de Araújo Nogueira – Vitória Virtual

Vitória Virtual
SOLUÇÕES EM DATA MINING

Recuperação da informação

- É uma ampla área da Ciência da Computação focada primeiramente em prover ao usuário o acesso a informações de interesse.

- Ela lida com a:
  - Representação;
  - Armazenamento;
  - Organização;
  - e acesso aos itens.

Recuperação da informação no passado

- Desde do momento em que a humanidade passou a registrar a informação, nós tivemos que aprender a organizar, armazenar e buscar essas informações. No início, a informação estava armazenada em:
  - Tabuletas de argila;
  - Rolos de papiro;
  - Livros.

- Devido essa necessidade, foram criadas as bibliotecas.

Recuperação da informação no passado

- As bibliotecas, na antiguidade, eram construções imponentes, no qual a maior parte do conhecimento de uma civilização estava armazenado nelas.

- Com o acesso restrito a poucas pessoas, devido ao número limitado de cópia dos documentos ou pela fragilidade dos mesmos.

Recuperação da informação no presente

- Hoje, nós mantemos as bibliotecas e trabalhamos com muito mais informação do que antigamente. Capacitamos pessoas para auxiliar o usuário no momento de busca.

- Só em 2008, a população dos Estados Unido visitou as bibliotecas 1,3 bilhões de vezes e checou mais de 2 bilhões de itens.

Recuperação da informação na computação

- Porém, no dias de hoje, a maior parte da informação está no mundo digital;
- A computação nos permitiu criar mais informações e compartilhar com qualquer pessoa;
- Assim gerando uma quantidade gigantesca de informação a todo momento;
- Devido a essa demanda nasceram as ferramentas de buscas.

Ferramentas de busca - Vantagens

- Agilizando o processo de indexação dos documentos;
- Possibilitando buscas para qualquer usuário;
- Atendimento simultâneo de usuários;
- Modernização nas bibliotecas;

Ferramentas de busca - Desvantagens

- Indexação manual;
- Número restrito de palavras-chave;
- Vocabulário específico para as chaves de busca;

Exemplo de acervo

- Um acervo com 1000 documentos;
- Assumindo que sejam necessários 15 segundo para indexar um documento;
- 250 minutos -> 4 horas de trabalho;

Exemplo de acervo – jornal A Tribuna

- O acervo do jornal A Tribuna disponibilizado online;
- Contém mais 300 mil páginas do jornal de 2003 a 2018;
- Assumimos que o acervo a ser indexado possui 300 mil documentos(páginas);
- E assumindo que sejam necessários 1 minuto para indexar um documento;
- **300 mil minutos -> 5000 horas -> 208 dias.**

Solução?

Solução!

Indexação automática

Indexação automática - Vantagens

- Uso de todo o conteúdo do documento como chave de busca;
- Indexação rápida;

Como funciona a indexação automática?

Como funciona?

- Leitura automática do conteúdo do documento:
  - OCR(Optical Character Recognition);
  - Metadados.

- O conteúdo é enviado ao sistema de busca e indexado em seguida;

E a classificação?

Classificação

- Pode ser feita automaticamente;
- Utilizando os dados da indexação;
- Com o supervisionamento do especialista humano;
- Diminuindo o esforço do especialista humano;

Existe uma ferramenta assim?

Existe uma ferramenta assim?

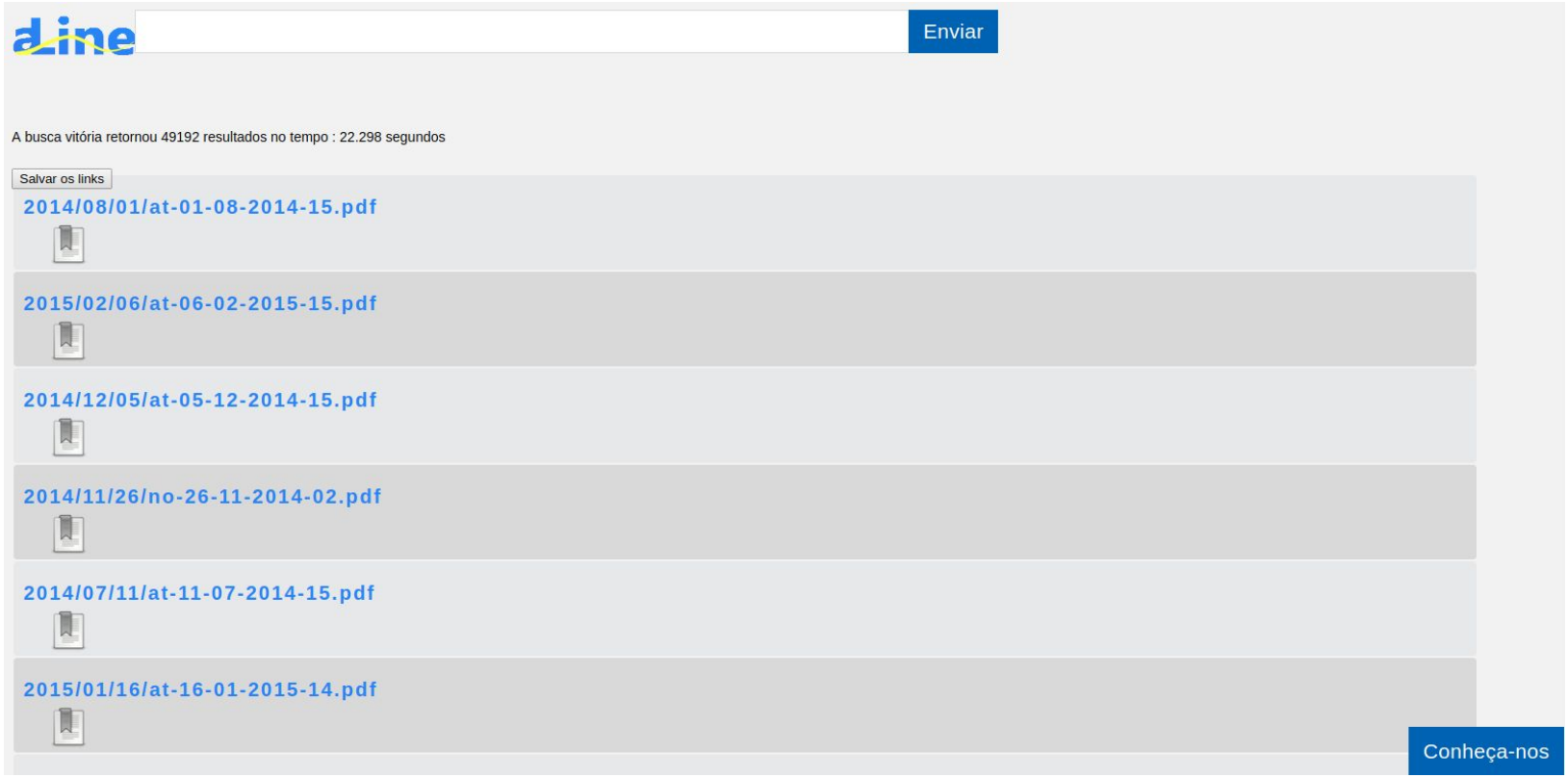
Sim

http://rii.lcad.inf.ufes.br/aline

aLine

- O aLine é buscador de conteúdo genérico.
- Além de prover resultados de busca, ele também auxilia na tarefa de classificação de um acervo e retorna alguns dados sobre o acervo.
- Atualmente trabalha com a base de dados do jornal A Tribuna.
- Aceita buscas por qualquer termo existentes nos documentos indexados e por metadados adicionados externamente;

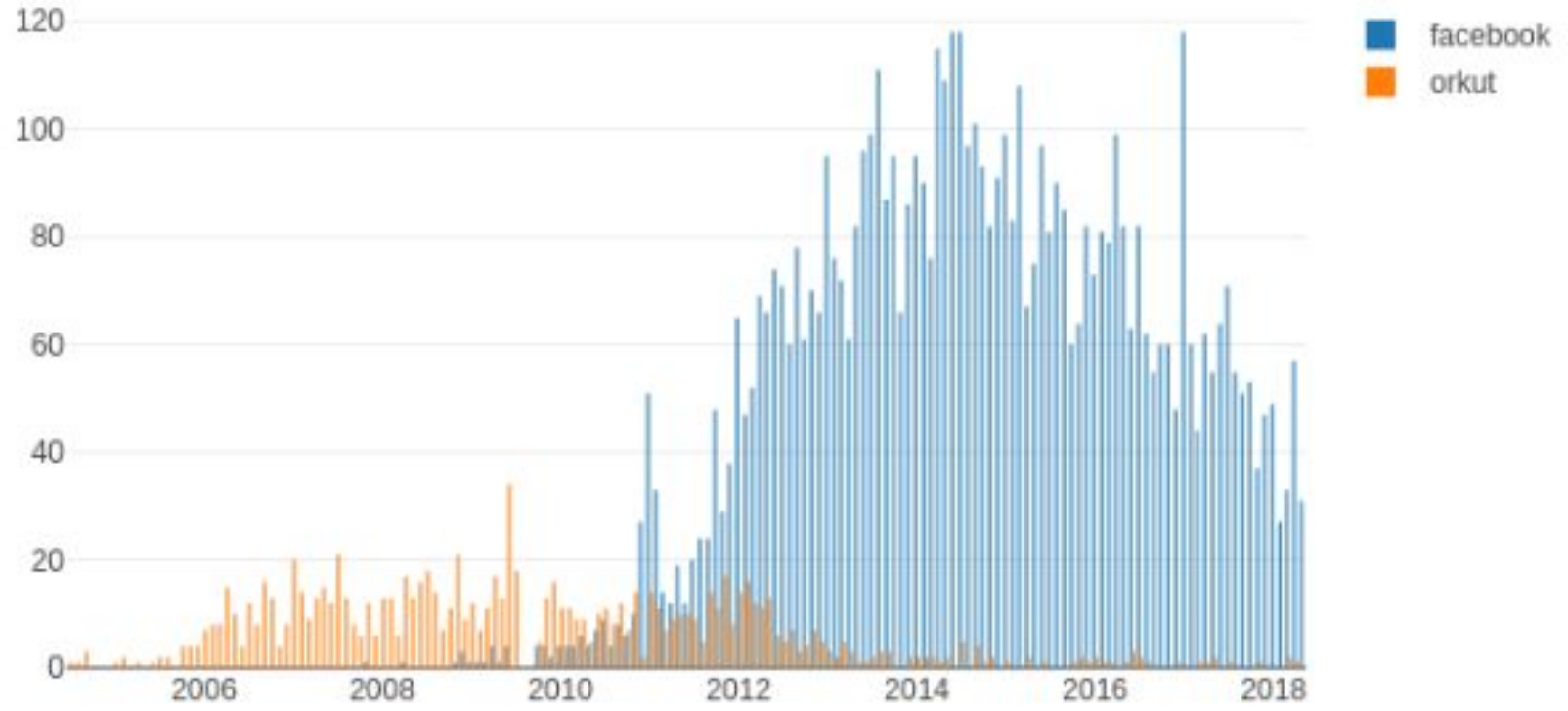
aLine


aLine
Enviar
A busca vitória retornou 49192 resultados no tempo : 22.298 segundos
Salvar os links
2014/08/01/at-01-08-2014-15.pdf
2015/02/06/at-06-02-2015-15.pdf
2014/12/05/at-05-12-2014-15.pdf
2014/11/26/no-26-11-2014-02.pdf
2014/07/11/at-11-07-2014-15.pdf
2015/01/16/at-16-01-2015-14.pdf
Conheça-nos

aLine


aLine
Enviar
Distribuição dos termos: nvidia, amd.
6
5
4
3
2
1
0
01/2014
02/2014
04/2014
07/2014
10/2014
12/2014
10/2015
11/2015
12/2015
01/2016
03/2016
07/2016
10/2016
08/2014
11/2014
01/2015
02/2015
07/2015
06/2016
nvidia
amd
Salvar os links
Conheça-nos

aLine


120
100
80
60
40
20
0
2006
2008
2010
2012
2014
2016
2018
facebook
orkut

Obrigado!

Matheus de Araújo Nogueira

Sócio Fundador

manogueira@vitoriavirtual.com.br

www.vitoriavirtual.com.br

TECVITÓRIA Incubadora

Vitória Virtual
SOLUÇÕES EM DATA MINING

UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
DOCETE OMNES GENTES
UFES