

# Análise do uso de *feedback* de relevância no Sistema de Integração Lattes-Qualis (SILQ)

Carlos Bonetti<sup>1</sup>

<sup>1</sup>Bacharelando de Ciência da Computação  
Departamento de Informática e Estatística  
Centro Tecnológico  
Universidade Federal de Santa Catarina

Orientação: Prof<sup>a</sup>. Dr<sup>a</sup>. Carina F. Dorneles

Trabalho de Conclusão de Curso, 2016/2

# Sumário

## Introdução

Histórico e Justificativa

## Conceitos

Information Retrieval e Data-matching

Métricas e avaliação de sistemas IR

## Objetivos

## Desenvolvimento

Alterações tecnológicas

Uso de feedback de relevância

## Conclusões

# Sumário

## Introdução

Histórico e Justificativa

## Conceitos

Information Retrieval e Data-matching

Métricas e avaliação de sistemas IR

## Objetivos


## Desenvolvimento

Alterações tecnológicas

Uso de feedback de relevância

## Conclusões

# Lattes



Dados Gerais	Linhas de pesquisa	Projetos	Áreas	Produção em C.T. & A.	Bancas	Eventos	Orientações	Citações
<div>  <div> <p><b>André Azevedo da Fonseca</b></p> <p>Coordenador do curso de Comunicação Social da Universidade de Uberaba. Pós-doutorando em Estudos Culturais no Programa Avançado de Cultura Contemporânea - PACCUPRJ, Doutor em História pela Universidade Estadual Paulista - Julio Mesquita Filho (Unesp). Especialista em História do Brasil pela PUC-MG. Graduado em Comunicação Social (Jornalismo) pela Universidade de Uberaba (Unube). Autor do livro-reportagem "Cotidianos culturais e outras histórias: a cidade sob novas olhares" (Unube, 2004) e de quatro capítulos de livros. Como estudante universitário conquistou 8 prêmios nacionais. Como professor orientador, conquistou 19 prêmios nacionais e internacionais. Em 2010 foi contemplado no Prêmio Rumos Pesquisa Aplicada do Itaú Cultural. Tem experiência na área de Livro-reportagem, Docência Superior e pesquisa acadêmica nas áreas de Comunicação, Educação e História. Os interesses de pesquisa, com ênfase na interdisciplinaridade, englobam Comunicação, Cultura e Hegemonia, Movimentos Sociais e Culturais Contemporâneos, Indústria criativa, Estudos Culturais, Cultura Política, Política Cultural, Direitos Culturais, Imaginários Sociais e Políticos, Mitologias Políticas, Redes sociais, Comunicação e Educação, Comunicação e Cidadania. (Texto informado pelo autor)</p> <p>Última atualização do currículo em 26/01/2012</p> <p>Endereço para acessar este CV: <a href="http://lattes.cnpq.br/5565508482791214">http://lattes.cnpq.br/5565508482791214</a></p> </div> <div>  <p><b>Certificado pelo autor em 26/01/12</b></p> </div> <div> <p><a href="#">Rede de colaboração</a></p> <p><a href="#">Rede de grupos de pesquisa</a></p> <p><a href="#">SILQ - artigos em texto completo</a></p> </div> </div>								
<p><b>Dados pessoais</b></p> <p><b>Nome</b> André Azevedo da Fonseca</p> <p><b>Nome em citações bibliográficas</b> FONSECA, A. A.</p> <p><b>Sexo</b> Masculino</p> <p><b>Endereço profissional</b> Universidade de Uberaba, Câmara de Ensino de Graduação, Pós-Graduação, Pesquisa e Extensão, Curso de Comunicação Social Av. Nere Sabino, 1801 - Bloco L Universitário 36055-500 - Uberaba, MG - Brasil Telefone: (34) 33109952 URL da Homepage: <a href="http://www.comunicacao.unube.br">http://www.comunicacao.unube.br</a></p>								
<p><b>Formação acadêmica/Titulação</b></p> <p><b>2011</b> Pós-Doutorado - Universidade Federal do Rio de Janeiro, UFRJ, Brasil Bolsista do(a) Itaú Cultural</p>								

# Qualis

ISSN	Título	Área de Avaliação	Estrato
1041-4347	IEEE Transactions on Knowledge and Data Engineering (Print)	CIÊNCIA DA COMPUTAÇÃO	►A1
0018-9464	IEEE Transactions on Magnetics	CIÊNCIA DA COMPUTAÇÃO	►B4
0278-0062	IEEE Transactions on Medical Imaging (Print)	CIÊNCIA DA COMPUTAÇÃO	►A1
1536-1233	IEEE Transactions on Mobile Computing	CIÊNCIA DA COMPUTAÇÃO	►A2
1520-9210	IEEE Transactions on Multimedia	CIÊNCIA DA COMPUTAÇÃO	►A2
2162-237X	IEEE Transactions on Neural Networks and Learning Systems	CIÊNCIA DA COMPUTAÇÃO	►A1
0018-9499	IEEE Transactions on Nuclear Science	CIÊNCIA DA COMPUTAÇÃO	►B1
1045-9219	IEEE Transactions on Parallel and Distributed Systems (Print)	CIÊNCIA DA COMPUTAÇÃO	►A2
0885-8950	IEEE Transactions on Power Systems	CIÊNCIA DA COMPUTAÇÃO	►B2
0098-5589	IEEE Transactions on Software Engineering	CIÊNCIA DA COMPUTAÇÃO	►A1
1083-4427	IEEE Transactions on Systems, Man and Cybernetics. Part A, S	CIÊNCIA DA COMPUTAÇÃO	►A2
1094-6977	IEEE Transactions on Systems, Man and Cybernetics. Part C, Ap	CIÊNCIA DA COMPUTAÇÃO	►A2
0018-9545	IEEE Transactions on Vehicular Technology	CIÊNCIA DA COMPUTAÇÃO	►A1
1063-8210	IEEE Transactions on Very Large Scale Integration (VLSI) System	CIÊNCIA DA COMPUTAÇÃO	►A2
1077-2626	IEEE Transactions on Visualization and Computer Graphics	CIÊNCIA DA COMPUTAÇÃO	►A2
1536-1276	IEEE Transactions on Wireless Communications	CIÊNCIA DA COMPUTAÇÃO	►A1
1536-1284	IEEE Wireless Communications	CIÊNCIA DA COMPUTAÇÃO	►A1
2162-2337	IEEE Wireless Communications Letters	CIÊNCIA DA COMPUTAÇÃO	►B4
1932-4537	IEEE eTransactions on Network and Service Management	CIÊNCIA DA COMPUTAÇÃO	►B3
1932-8540	IEEE-RITA	CIÊNCIA DA COMPUTAÇÃO	►B5
1545-5963	IEEE/ACM Transactions on Computational Biology and Bioinform	CIÊNCIA DA COMPUTAÇÃO	►B1
0916-8532	IEICE Transactions on Information and Systems	CIÊNCIA DA COMPUTAÇÃO	►B1
1751-861X	IET Computers & Digital Techniques (Online)	CIÊNCIA DA COMPUTAÇÃO	►B1
1751-8601	IET Computers & Digital Techniques (Print)	CIÊNCIA DA COMPUTAÇÃO	►B1
1751-8806	IET Software (Print)	CIÊNCIA DA COMPUTAÇÃO	►B1
1091-9856	INFORMS Journal on Computing	CIÊNCIA DA COMPUTAÇÃO	►B1
1526-5528	INFORMS Journal on Computing (Online)	CIÊNCIA DA COMPUTAÇÃO	►B1

## Histórico e Justificativa

- ▶ AGUIAR, Felipe Nedel de; COSTA, Maria Eloísa. **SILQ - Sistema de Integração Lattes Qualis**. Trabalho de Conclusão de Curso. Florianópolis: Universidade Federal de Santa Catarina, Biblioteca Universitária, 2015.
- ▶ Qualificação automática de produções científicas através de busca por similaridade textual nos dados Qualis;

Home Participe da nossa pesquisa de usabilidade carlosbonetti.mail@gmail.com Sobre Contato Logout **SILQ**

Sistema de Integração Lattes-Qualis

Meus dados

Grupos de avaliação **1**

Avaliação livre

## Resultado da avaliação

Nome: Carina Friedrich Dorneles

Área do conhecimento: Ciência da Computação

Sub área do conhecimento: Metodologia e Técnicas da Computação

Grande área do conhecimento: Ciências Exatas e da Terra

Especialidade: Banco de Dados

Totalizador (para nível de confiança Normal): 4x A1 - 1x A2 - 5x B1 - 3x B2 - 6x B3 - 1x B4 - 4x B5 - 1x C

Área utilizada na avaliação: Ciência da Computação

## Artigos


Título do Artigo	Ano de publicação	Título Periódico ou Revista	ISSN	Conceito
Web table taxonomy and formalization	2013	SIGMOD Record	0163-5808	A1 (1.0) 

Figura: Primeira versão do SILQ (<http://silq.inf.ufsc.br>)

# Sumário

## Introdução

Histórico e Justificativa

## Conceitos

Information Retrieval e Data-matching

Métricas e avaliação de sistemas IR

## Objetivos

## Desenvolvimento

Alterações tecnológicas

Uso de feedback de relevância

## Conclusões



# IR e Data Matching

- ▶ *Information Retrieval (IR)*
  - ▶ *query*
  - ▶ *documentos*
- ▶ *Data-Matching*
  - ▶ *similaridade / dissimilaridade*
  - ▶ *threshold*

## *n*-grams / trigrams

Revista:  $A = \{\_R, \_Re, Rev, evi, vis, ist, sta, ta\_ \}$

Revisor:  $B = \{\_R, \_Re, Rev, evi, vis, iso, sor, or\_ \}$

5 elementos em comum:  $|A \cap B|$

11 elementos distintos:  $|A \cup B|$

$$\text{trigrams}(\text{Revista}, \text{Revisão}) = \frac{|A \cap B|}{|A \cup B|} = \frac{5}{11} = 0.45 = 45\%$$

- ▶ SILQ: sistema de IR baseado em *data matching*
- ▶ Utiliza trigrams para *matching* entre eventos informados no Lattes e os registrados no Qualis
- ▶ *Threshold* de 0.6 ('nível de confiança normal')

## Como o SILQ avalia um currículo Lattes

Artigo #1 (extraído do Lattes)

**Título:** Approximate data instance matching: a survey

**Ano:** 2011

**Área:** Ciência da Computação

**Journal:** Knowledge and Information Systems

**ISSN:** 0219-1377

Artigo #2 (extraído do Lattes)

...

## Como o SILQ avalia um currículo Lattes

### Artigo #1

**Título:** Approximate data instance matching: a survey

**Ano:** 2011

**Área:** Ciência da Computação

**Journal:** Knowledge and Information Systems

**ISSN:** 0219-1377

query: (ISSN, ano, área)

$q_A = (0219-1377, 2011, \text{Ciência da Computação})$

$q_A = (0219-1377, 2011, \text{Ciência da Computação})$

$q_A = (0219-1377, 2011, \text{Ciência da Computação})$

Conceito	Ano	ISSN	Título
A2	<b>2011</b>	0219-1377	Knowledge and Information Systems (Print)
A2	2012	0219-1377	Knowledge and Information Systems (Print)
B1	2014	0219-1377	Knowledge and Information Systems (Print)
A2	2010	0219-1377	Knowledge and Information Systems (Print)

**Tabela:** Resultados retornados pelo SILQ para a query  $q_A$

$q_A = (0219-1377, 2011, \text{Ciência da Computação})$

Conceito	Ano	ISSN	Título
A2	<b>2011</b>	0219-1377	Knowledge and Information Systems (Print)
A2	2012	0219-1377	Knowledge and Information Systems (Print)
B1	2014	0219-1377	Knowledge and Information Systems (Print)
A2	2010	0219-1377	Knowledge and Information Systems (Print)

**Tabela:** Resultados retornados pelo SILQ para a query  $q_A$

## Resultado

Artigo #1 recebe o conceito A2



## Como o SILQ avalia um currículo Lattes

Trabalho #1 (extraído do Lattes)

**Título:** A Strategy for Allowing Meaningful and Comparable Scores in Approximate Matching

**Ano:** 2007

**Área:** Ciência da Computação

**Evento:** Conference on Information and Knowledge Management (CIKM)

## Como o SILQ avalia um currículo Lattes

Trabalho #1 (extraído do Lattes)

**Título:** A Strategy for Allowing Meaningful and Comparable Scores in Approximate Matching

**Ano:** 2007

**Área:** Ciência da Computação

**Evento:** Conference on Information and Knowledge Management (CIKM)

query: (título do evento, ano, área)

$q_T = (\text{Conference on Information and Knowledge Management (CIKM), 2007, Ciência da Computação})$

$q_T =$  (Conference on Information and Knowledge  
Management (CIKM), 2007, Ciência da Computação)

$q_T =$  (Conference on Information and Knowledge Management (CIKM), 2007, Ciência da Computação)

Conceito	Similaridade	Título
A1	0.71	International Conference on Information and Knowledge Management
B4	0.64	International Conference on Information, Process, and Knowledge Management

**Tabela:** Resultados retornados pelo SILQ para a query  $q_T$

$q_T$  = (Conference on Information and Knowledge Management (CIKM), 2007, Ciência da Computação)

Conceito	Similaridade	Título
A1	0.71	International Conference on Information and Knowledge Management
B4	0.64	International Conference on Information, Process, and Knowledge Management

**Tabela:** Resultados retornados pelo SILQ para a query  $q_T$

## Resultado

Trabalho #1 recebe o conceito A1

# Motivação

- ▶ Qual o *threshold* ideal para o SILQ?
- ▶ Qual a taxa de acerto do sistema? Ele está avaliando corretamente os currículos Lattes?
- ▶ É possível aumentar a taxa de acerto utilizando *feedback* de usuários?
- ▶ Atualização tecnológica e da base de dados

# Métricas e avaliação de sistemas de IR (TODO)

- ▶ Taxa de acerto / Exatidão
- ▶ Falar de Precisão e Revocação?
- ▶ Falar de Precision at k e R-Precision ?
- ▶ Média de rank Recíproco
- ▶ Conjunto de Testes

# Feedback de relevância

TODO: breve de explicação de feedback de relevância



# Sumário

## Introdução

Histórico e Justificativa

## Conceitos

Information Retrieval e Data-matching

Métricas e avaliação de sistemas IR

## Objetivos

## Desenvolvimento

Alterações tecnológicas

Uso de feedback de relevância

## Conclusões

# Objetivos

## Objetivo geral

Analisar o impacto que o uso de feedback de relevância tem na precisão dos resultados de avaliações realizadas pelo SILQ, efetuado sobre uma nova arquitetura da ferramenta que inclui a criação de API de integração com outros sistemas e a atualização da base de dados conforme as novas classificações Qualis.

## Objetivos específicos

1. Reestruturação da arquitetura e banco de dados do SILQ a fim de suportar classificações de eventos e periódicos disponibilizados em um ritmo anual;

## Objetivos específicos

1. Reestruturação da arquitetura e banco de dados do SILQ a fim de suportar classificações de eventos e periódicos disponibilizados em um ritmo anual;
2. Atualização do banco de dados do sistema com as últimas classificações disponibilizadas pelo Qualis (anos 2013 e 2014);

## Objetivos específicos

1. Reestruturação da arquitetura e banco de dados do SILQ a fim de suportar classificações de eventos e periódicos disponibilizados em um ritmo anual;
2. Atualização do banco de dados do sistema com as últimas classificações disponibilizadas pelo Qualis (anos 2013 e 2014);
3. Criação de uma API pública de disponibilização dos serviços do SILQ, via camada de aplicação REST para integração com outros sistemas;

## Objetivos específicos

4. Alterações na interface do sistema incluindo migração de *framework* de interface, inclusão de controles de *feedback*, novos gráficos de acompanhamento de grupos de pesquisa e melhorias gerais de usabilidade;

## Objetivos específicos

4. Alterações na interface do sistema incluindo migração de *framework* de interface, inclusão de controles de *feedback*, novos gráficos de acompanhamento de grupos de pesquisa e melhorias gerais de usabilidade;
5. Propor novos algoritmos de avaliação baseados em similaridade textual e *feedback* de relevância e verificar se a taxa de acerto do sistema foi melhorada com tal ação.

# Sumário

## Introdução

Histórico e Justificativa

## Conceitos

Information Retrieval e Data-matching

Métricas e avaliação de sistemas IR

## Objetivos

## Desenvolvimento

Alterações tecnológicas

Uso de feedback de relevância

## Conclusões



# Extração e inserção dos novos dados Qualis

- ▶ Até final de 2015:
  - ▶ Qualis trienal
  - ▶ 2010-2012
  - ▶ PDFs e planilhas XLS
- ▶ Início de 2016
  - ▶ Qualis anual
  - ▶ 2010, 2011, 2012, 2013, 2014
  - ▶ Planilhas CSV
- ▶ Limpeza manual (erros de codificação, ISSNs omitidos, etc.)
- ▶ 339.204 registros

# Atualização tecnológica

- ▶ Criação da camada REST de integração
- ▶ Migração de *framework*: *Play* → *Spring*
- ▶ Reescrita do *front-end* com *AngularJS*
  - ▶ Novos gráficos de avaliação para grupos de pesquisa
  - ▶ Melhorias no módulo de usuários
  - ▶ Redesign da página de resultados de avaliação
  - ▶ Inclusão dos controles de *feedback*
- ▶ Garantida da qualidade com testes automatizados

# Obtenção de feedback

2016 | Ajustamento de pesos para ratings de múltiplos critérios em recomendação de itens  
Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia)

B3 23% 2012 WEBMEDIA Brazilian Symposium on Multimedia and the Web

2015 | Towards Automatic Document Classification by Exploiting only Knowledge Resources  
International Conference of the Chilean Computer Science Society

B3 100% 2012 SCCC International Conference of the Chilean Computer Science Society

B2 64% 2012 ICSC\_A International Computer Science Conference

[Ver menos resultados](#)

2015 | Implementação de um esquema de extração de dados tabulares da web  
XII Workshop de Trabalhos de Iniciação Científica (WTIC)

Nenhum registro Qualis correspondente

Nenhum conceito encontrado | [Sugerir matching](#)

Sugerir matching

Nenhum registro Qualis correspondente

**Figura:** Controles de feedback da página de resultados de avaliação do SILQ

# Algoritmo $fb(t)$

TODO: Explicação do algoritmo + exemplo

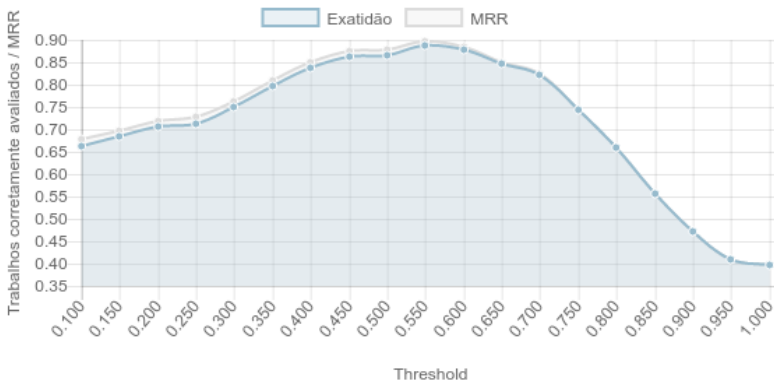
# Algoritmo query\_aliasing

TODO: Explicação do algoritmo + exemplo

# Avaliação experimental

- ▶ Conjunto de testes
  - ▶ 33 currículos de pesquisadores do PPGCC
  - ▶ 300 trabalhos aleatoriamente selecionados e avaliados manualmente
- ▶ Comparação entre o resultado retornado pelo sistema e o resultado selecionado

## Avaliação de *threshold* ideal



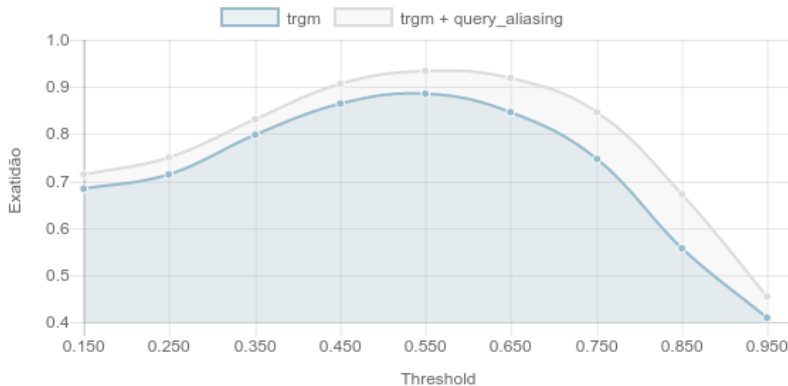
**Figura:** Valores de exatidão e MRR para diferentes valores de *threshold* utilizando o método *trigram*

# Exatidão dos algoritmos propostos

Algoritmo	Exatidão
<i>trgm</i>	88.667%
<i>trgm + fb(1.00)</i>	89.667%
<i>trgm + fb(0.90)</i>	90.667%
<i>trgm + fb(0.80)</i>	92.667%
<i>trgm + fb(0.70)</i>	92.667%
<i>trgm + fb(0.60)</i>	91.000%
<i>trgm + query_aliasing</i>	<b>93.333%</b>

**Tabela:** Comparação da exatidão dos diferentes algoritmos testados (utilizando *threshold* de 0.55)





**Figura:** Comparação da taxa de acerto do algoritmo *trgm* e do *trgm + query\_aliasing* para diferentes *thresholds*

# Sumário

## Introdução

Histórico e Justificativa

## Conceitos

Information Retrieval e Data-matching

Métricas e avaliação de sistemas IR

## Objetivos

## Desenvolvimento

Alterações tecnológicas

Uso de feedback de relevância

## Conclusões

# Conclusões

- ▶ Criação da camada REST de integração
  - ▶ Ex.: `http://silq.inf.ufsc.br/api/qualis`

# Conclusões

- ▶ Criação da camada REST de integração
  - ▶ Ex.: `http://silq.inf.ufsc.br/api/qualis`
- ▶ Atualização da base de dados com os novos registros Qualis

# Conclusões

- ▶ Criação da camada REST de integração
  - ▶ Ex.: `http://silq.inf.ufsc.br/api/qualis`
- ▶ Atualização da base de dados com os novos registros Qualis
- ▶ Métricas de exatidão do sistema

# Conclusões

- ▶ Criação da camada REST de integração
  - ▶ Ex.: `http://silq.inf.ufsc.br/api/qualis`
- ▶ Atualização da base de dados com os novos registros Qualis
- ▶ Métricas de exatidão do sistema
- ▶ Descoberto *threshold* ideal: 0.55

# Conclusões

- ▶ Criação da camada REST de integração
  - ▶ Ex.: `http://silq.inf.ufsc.br/api/qualis`
- ▶ Atualização da base de dados com os novos registros Qualis
- ▶ Métricas de exatidão do sistema
- ▶ Descoberto *threshold* ideal: 0.55
- ▶ Inserção dos controles de *feedback* de relevância

# Conclusões

- ▶ Criação da camada REST de integração
  - ▶ Ex.: `http://silq.inf.ufsc.br/api/qualis`
- ▶ Atualização da base de dados com os novos registros Qualis
- ▶ Métricas de exatidão do sistema
- ▶ Descoberto *threshold* ideal: 0.55
- ▶ Inserção dos controles de *feedback* de relevância
- ▶ Taxa de acerto média do sistema melhorada de 87% para 93.3% com o uso de *feedback* de usuários



## Trabalhos futuros

- ▶ Avaliar outras funções de similaridade

## Trabalhos futuros

- ▶ Avaliar outras funções de similaridade
- ▶ Avaliar diferentes estratégias de uso de *feedback* de relevância
  - ▶ Ex.: Algoritmo de Rocchio, *machine learning*, etc

## Trabalhos futuros

- ▶ Avaliar outras funções de similaridade
- ▶ Avaliar diferentes estratégias de uso de *feedback* de relevância
  - ▶ Ex.: Algoritmo de Rocchio, *machine learning*, etc
- ▶ Tradução de nomes de eventos

## Trabalhos futuros

- ▶ Avaliar outras funções de similaridade
- ▶ Avaliar diferentes estratégias de uso de *feedback* de relevância
  - ▶ Ex.: Algoritmo de Rocchio, *machine learning*, etc
- ▶ Tradução de nomes de eventos
- ▶ Automatizar ainda mais o processo de avaliação de Programas de Pós-Graduação conforme regras da CAPES
  - ▶ Gerar valores de  $I_{geral}$  e  $I_{restrito}$
  - ▶ Utilizar pesos considerados pela CAPES

## Análise do uso de *feedback* de relevância no Sistema de Integração Lattes-Qualis (SILQ)

Dúvidas?

Carlos Bonetti

*carlosbonetti.mail@gmail.com*