

Programa:	Engenharia Elétrica
Área de Concentração:	Engenharia de Computação
Aluno:	Jefferson Carlos de Mendonça
Orientador:	Prof. Dr. Edson Satoshi Gomi
Curso:	Mestrado
Data de Ingresso:	14/09/2016
Título:	Algoritmo para categorização das Perguntas do <i>Site Stack Overflow</i>

## Resumo

Transformações no cenário tecnológico provocam mudanças constantes e exigem atualização contínua. Para manter-se atualizado, profissionais da área de computação recorrem à diversas fontes de informação, dentre elas destaca-se o site *Stack Overflow*[1], maior comunidade de perguntas e respostas, onde os usuários podem aprender, trocar experiências e compartilhar conhecimento. O objeto de pesquisa deste trabalho é propor um algoritmo que consiga categorizar as perguntas deste *website*. Uma vez que os questionamentos estejam devidamente indexados e organizados em tópicos, será possível detectar as dúvidas mais frequentes reportadas pelos usuários, permitindo que um assunto seja encontrado com maior facilidade, além de permitir que as entidades detentoras das tecnologias envolvidas, use os resultados obtidos para criar manuais e cursos que potencialize o entendimento da comunidade de interesse.

**Palavras-chave:** recuperação de informação, mineração de textos, classificação de textos, categorização automática de textos, extração de palavras chave, indexação de documentos, stack overflow.

# Algoritmo para categorização das Perguntas do *Site Stack Overflow*

## 1 Introdução

A evolução na área computacional se desenvolve em ritmo acelerado, algoritmos, técnicas para programação distribuída, testes automatizados, bancos de dados e outros assuntos possuem algo em comum, eles mudam com frequência. Com as *linguagens de programação* não é diferente, suas API's são aprimoradas constantemente e muitas vezes elas incorporam novos paradigmas. Para acompanhar estas mudanças, profissionais da área de computação necessitam qualificação, que pode ser obtida através de cursos presenciais, a distância, livros, revistas, artigos e claro *websites*.

Manning et al. (2009) afirmam que a *web* tornou-se a principal fonte por busca de informação e o relatório elaborado por Fallows (2004) conclui que « 92% dos internautas dizem que a internet é um bom lugar para obter informações diariamente ». Dentre estas fontes, merece destaque o fórum *Stack Overflow* (SO), maior comunidade *online* de perguntas e respostas onde os usuários podem aprender, trocar experiências e compartilhar conhecimento.

### 1.1 Contextualização do Problema

O número de questões em sites de perguntas e respostas cresce diariamente, faz-se necessário categorizar os assuntos discutidos em tópicos, para uma busca mais eficiente e rápida. Manning et al. (2009) comparam este problema ao da procura de um livro em uma biblioteca, com certeza nossa busca será mais rápida e assertiva se os livros estiverem separados em prateleiras por assunto ou tópico. Yasotha and Charles (2016) adicionam que a categorização manual de textos pode ser feita somente por especialistas e essa tarefa requer muito tempo. Como consequência é de grande importância a categorização e classificação de documentos de forma automática, ajudando os usuários a encontrarem informações relevantes para as suas necessidades.

### 1.2 Objetivos

A proposta deste projeto de pesquisa, é desenvolver um algoritmo que categorize e classifique os assuntos discutidos no SO, que faça uso apenas do texto disponível nas perguntas e respostas, sem que haja a necessidade de recorrer ao uso de tags.

### 1.3 Justificativas

O método muito bem detalhado por Arash et al. (2016) categoriza os dados do SO. Em seu projeto os autores classificaram os assuntos utilizando as tags da própria pergunta. No SO o próprio usuário, elege qual é a tag da referida pergunta publicada. Apesar do

grande avanço nesse campo de pesquisa, o uso de tags limita a expansão da solução para outros fóruns que não possuem este recurso, como por exemplo os *sites* Quora[2] e Code Ranch[3].

## 1.4 Organização do texto

Para melhor definir qual o posicionamento do presente projeto, no capítulo seguinte será detalhado em maior profundidade os projetos que abordaram a categorização de documentos, inclusive àqueles que também fizeram uso do site *Stack Overflow* como base de dados. Então a proposta será detalhada quanto aos procedimentos para a indexação das perguntas e respostas, desde a seleção do conteúdo original, armazenamento em banco de dados e extração, por fim serão exibidos os resultados esperados. Segundo Kaleta (2014) o termo indexação pode ser entendido como um dicionário de palavras-chave que representam o conteúdo de um texto.

## 2 Revisão da Literatura

Revisão da literatura (item pendente).

## 3 Detalhamento da Proposta

Dados de 2008 apontam que o SO recebe cerca de 7.000 questões por dia, Krippendorff (2012) sublinha que a mineração de texto vem sendo adotada como uma alternativa à situações em que a análise manual de grande quantidade de dados se torna inviável. Análogo a metodologia proposta por Arash et al. (2016) a figura 1 exemplifica o processamento da informação desde sua obtenção até a produção dos resultados finais.

A primeira etapa é composta pela **coleta da informação**, o SO faz parte da rede de *websites* StackExchange[4] que adota uma política de acesso livre à informação, ou seja, as postagens - perguntas e respostas, os usuários e os comentários são periodicamente disponibilizados de forma gratuita[5] para download e análise.

Na etapa de **extração da informação** será desenvolvido um algoritmo para extrair os assuntos principais Turney (2000) das dúvidas postadas no *site*, assim como na proposta anterior também será utilizado o Wikipédia Miner Milne and Witten (2012) como base de conhecimento sobre o vocabulário (palavras-chave) encontrado, de forma complementar também serão armazenados a quantidade de visualizações, a pontuação atribuída e o número de respostas da referida pergunta.

Por fim, na fase de **mineração do texto** será criado um *ranking* Mihalcea and Tarau (2004a) dos tópicos relevantes questionados com mais frequência, será utilizado o Gephi Bastian et al. (2009) software *open source* para exibir o grafo das categorias e suas interconexões.

## 4 Plano de Trabalho

### 4.1 Resultados Desejados e Validação

A análise será feita sobre os mesmos dados da pesquisa realizada por Arash et al. (2016), os resultados obtidos anteriormente serão a base para a medição da acurácia da nova

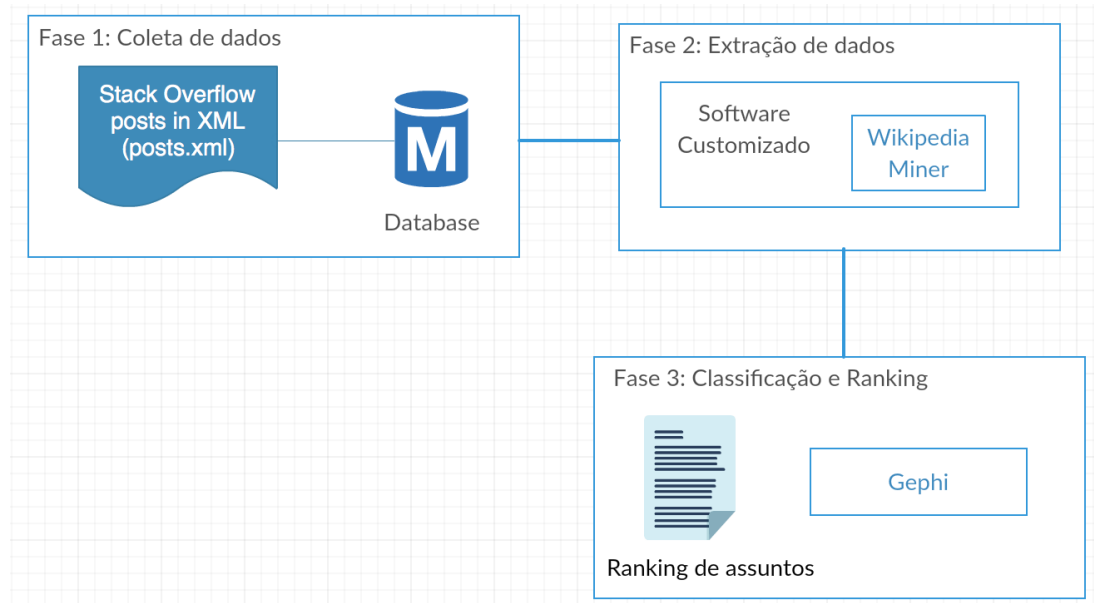


Figura 1: Visão geral para categorização de perguntas e resposta do Stack Overflow.

proposta e então será possível aplicar o modelo desenvolvido sem a utilização de tags pré-definidas, possibilitando a construção de catálogos para outros sites de perguntas e respostas.

## 4.2 Atividades e Cronograma

Atividades:

- Disciplinas
  - Machine Learning
  - Metodologia Científica
  - Ciência de Dados
  - Inteligência Artificial
  - Fundamentos Engenharia de Computação
- Revisão Literária
- Plano de Pesquisa
- Artigo Científico
- Projeto de Pesquisa
  - Fase 1: Coleta de dados
  - Fase 2: Extração de dados
  - Fase 3: Classificação e *Ranking*
- Qualificação
- Defesa

Etapas / Tempo	Cronograma do Programa de Pós Graduação - Mestrado																																																							
	2016												2017												2018												2019																			
	SET	OUT	NOV	DEZ	JAN	FEV	MAR	ABR	MAI	JUN	JUL	AGO	SET	OUT	NOV	DEZ	JAN	FEV	MAR	ABR	MAI	JUN	JUL	AGO	SET	OUT	NOV	DEZ	JAN	FEV	MAR	ABR	MAI	JUN	JUL	AGO	SET	OUT	NOV	DEZ	JAN	FEV	MAR	ABR	MAI	JUN	JUL	AGO	SET	OUT	NOV	DEZ				
DISCIPLINAS																																																								
Machine L.																																																								
Metodologia C.																																																								
Clência Dados																																																								
Inteligência A.																																																								
Fund. Eng. Comp.																																																								
Revisão Lit/Plano																																																								
Revisão Literária																																																								
Plano Pesquisa																																																								
Artigo Científico																																																								
Projeto Pesquisa																																																								
Fase 1																																																								
Fase 2																																																								
Qualificação																																																								
Fase 3																																																								
Defesa																																																								

Figura 2: Cronograma.

## Notas

1. <http://stackoverflow.com/tour>
2. <http://www.quora.com>
3. <http://coderanch.com>
4. <http://stackexchange.com>
5. <https://archive.org/download/stackexchange>

## Referências

- Arash, M., English, E., and Mahdi (2016). Text mining stackoverflow An insight into challenges and subject-related difficulties faced An insight into challenges and subject-related difficulties faced by computer science learners subject-related difficulties faced by computer science learners. *Journal of Enterprise Information Management Journal of Enterprise Information Management Journal of Enterprise Information Management Web of Science Database Library Review*, 29(3):255–275.
- Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. *International AAAI Conference on Web and Social Media Third International AAAI Conference on Weblogs and Social Media*.
- Fallows, D. (2004). The Internet and Daily Life many americans use the internet in everyday activities, but traditional offline habits still dominate. [http://www.pewinternet.org/files/old-media/Files/Reports/2004/PIP\\_Internet\\_and\\_Daily\\_Life.pdf](http://www.pewinternet.org/files/old-media/Files/Reports/2004/PIP_Internet_and_Daily_Life.pdf). Accessed: 2016-11-08.
- Joorabchi, A., English, M., and Mahdi, A. E. (2015). Automatic mapping of user tags to Wikipedia concepts: The case of a Q&A website – StackOverflow. *Article Journal of Information Science*, 41(5):570–583.
- Kaleta, Z. (2014). Semantic text indexing. *Computer Science*, 15(1).
- Krippendorff, K. (2012). *Content Analysis: An Introduction to Its Methodology*. SAGE Publications, Inc, Thousand Oaks, CA, 3 edition.
- Manning, C. D., Raghavan, P., and Schütze, H. (2009). An Introduction to Information Retrieval. *Online*, 1(c):569.

- Mihalcea, R. (2007). Using Wikipedia for Automatic Word Sense Disambiguation. In *Proceedings of NAACL HLT*, volume 2007, pages 142–147.
- Mihalcea, R. and Tarau, P. (2004a). Textrank: Bringing order into texts. In Lin, D. and Wu, D., editors, *Proceedings of EMNLP 2004*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.
- Mihalcea, R. and Tarau, P. (2004b). TextRank: Bringing order into texts. *Proceedings of EMNLP*, 85:404–411.
- Mihalcea, R. F. and Mihalcea, S. I. (2001). Word Semantics for Information Retrieval: Moving One Step Closer to the Semantic Web. *13th IEEE International Conference on Tools with Artificial Intelligence. ICTAI 2001*, pages 280–287.
- Milne, D. (2012). An open-source toolkit for mining wikipedia. In *In Proc. New Zealand Computer Science Research Student Conf*, page 2009.
- Milne, D. and Witten, I. H. (2012). An open-source toolkit for mining Wikipedia. *Artificial Intelligence*, 1:1–18.
- Miotto, R. and Weng, C. (2013). Unsupervised mining of frequent tags for clinical eligibility text indexing. *Journal of Biomedical Informatics*, 46(6):1145–1151.
- Posch, L. (2014). Enriching Ontologies with Encyclopedic Background Knowledge for Document Indexing. *Proceedings of the 13th International Semantic Web Conference*, pages 537–544.
- Roul, R. K., Asthana, S. R., and Sahay, S. K. (2015). Automated document indexing via intelligent hierarchical clustering: A novel approach. *2014 International Conference on High Performance Computing and Applications, ICHPCA 2014*.
- Turney, P. D. (2000). Learning algorithms for keyphrase extraction. *Inf. Retr.*, 2(4):303–336.
- Udell, J. (2005). UIMA and the Blogosphere. *1803072320050822*, page 30.
- Yasotha, R. and Charles, E. Y. A. (2016). Automated text document categorization. *2015 IEEE 7th International Conference on Intelligent Computing and Information Systems, ICICIS 2015*, pages 522–528.