Um estudo direcionado à comparação de resultados de algoritmos de extração de palavras-chave oriundas de artigos científicos em língua portuguesa

Glauco Primo e **Fernando Costa de Souza**
Engenharia de dados e do Conhecimento
Universidade Federal do Rio de Janeiro
gprimo@cos.ufrj.br, fernandocosta@cos.ufrj.br

Resumo

Palavras-chave são de suma importância no contexto de busca e recuperação de documentos. Cada vez torna-se mais importante conseguir buscar de maneira eficiente na enorme quantidade de informações que obtem-se na internet, principalmente em tempos de ensino remoto. Neste contexto, torna-se indispensável a utilização de palavras-chave em documentos científicos afim de agilizar a compreensão e revelância de determinado documento.

Existem diversos trabalhos em língua inglesa que descrevem bem o comportamento de algoritmos extratores de palavras-chave e o quão bem eles performam. Contudo, não existem muitos trabalhos que fazem esse tipo de relação qualitativa em língua portuguesa, principalmente devido a dificuldade de obter uma boa base de dados para certificar a competência do algoritmo.

Este trabalho visa justamente, levantar uma base de dados em língua portuguesa, e verificar o grau de acertos dos já conhecidos algoritmos de extração de palavras.

Palavras-chave

Palavras-chave; língua portuguesa; extratores de palavra; single rank;

1 Introdução

Existe uma fonte de informações rápida, fidedigna, eficaz e científica que pode ser obtida em mecanismos de busca de alta relevância: artigos acadêmicos. Porém ao buscar por um termo simples como por exemplo, Covid-19, nestes mecanismos, obtem-se quase 5 milhões de documentos.

Neste mar de documentos, relevantes ou não, para sua pesquisa, torna-se não só importante como essencial, o uso de palavras-chave adequadas. Algoritmos de extração de palavras-chave de textos tornam-se ferramentas valiosas para ajudar na obtensão dos principais termos que melhor resumem um documento.

Existem diversos trabalhos realizados em língua inglesa que demonstram o comportamento de algoritmos extratores de palavras-chave, tanto os supervisionados (Ian H. Witten *et al.*, 1999), como os não supervisionados(Kazi Saidul Hasan and Vincent Ng., 2010). Contudo, não é uma tarefa fácil encontrar uma boa base classificada em língua portuguesa para testar a eficácia de tais algoritmos.

Na abordagem supervisionada, as palavras-chave são classificadas de forma binária, ou seja, é sim uma palavra-chave, ou não é uma palavra-chave. Existem diversos algoritmos de aprenzidado de máquina possíveis, basta obter um dataset devidamente classificado para tal feito. Ian H. Witten propõe por exemplo, *KEA*, um sofisticado algoritmo que utiliza métodos léxicos com aprendizado de máquina (Ian H. Witten *et al.*, 1999) para predizer as palavras-chave.

Os algoritmos não supervisionados podem ser resumidos como um problema de classificação utilizando grafos (Kazi Saidul Hasan and Vincent Ng., 2010), onde os nós são os termos e as arestas

representam relações de recorrência. Basta obter, por exemplo, os 10 melhores nós e suas relações de recorrência no texto. Deve-se notar que as palavras que melhor sintentizam o texto, costumam aparecer logo no ínicio do documento quando se tratando de trabalhos acadêmicos. Logo, os primeiros termos encontrados, costumam recorrentemente serem os melhores, como é possível verificar na figura 1.

O estudo avaliou a prevalência de sintomas de depressão e ansiedade em uma amostra de trabalhadores brasileiros de diversos segmentos, durante a pandemia da Covid-19. Foi também verificada a correlação entre as escalas de ansiedade e depressão dos instrumentos de rastreio. Foram coletados dados on-line por meio de três instrumentos: questionário sociodemográfico e ocupacional, a Depression, Anxiety and Stress Scale - Short Form e o Inventário de Saúde Mental Ocupacional. Participaram 503 profissionais, destes 78,5% do sexo feminino, com idade média de 41,38 anos, das quais 92% cursaram o ensino superior e residiam na região Sul do Brasil. Ambas as escalas detectaram maior prevalência de sintomas de ansiedade em mulheres (54,3% e 59,9%) e em pessoas solteiras (68,8% e 68,1%). Houve associação significativa entre desfechos de sintomas de ansiedade e depressão e prevalência de duas variáveis independentes: o contato com pessoas diagnosticadas com Covid-19 e sentir-se preocupado com a pandemia. O Inventário de Saúde Mental Ocupacional mostrou maior sensibilidade para aferir sintomas de ansiedade e discriminar os trabalhadores que apresentam sintomas daqueles que indicam ter saúde mental, quando comparado ao outro instrumento. Sugerem-se estudos longitudinais para capturar os efeitos de longo termo dos desfechos avaliados, a fim de aperfeiçoar a análise dos preditores dos valores críticos e não críticos dos sintomas de agravos à saúde mental.

Palavras-chave escolhidas pelo autor:
trabalhadores; prevalência; ansiedade; depressão; covid-19

Figura 1: Resumo retirado do *abstract* de um dos artigos do conjunto de dados, seguido das palavras-chave escolhidas pelo autor (Romilda Guilland *et al.*, 2021)

O intuito deste artigo é verificar se os algoritmos extratores de palavras-chave não supervisionados como Single Rank (Xiaojun Wan e Jianguo Xiao [s.d.]) e Position Rank (Corina Florescu e Cornelia Caragea [s.d.]) tem uma performance melhor ou pior que na literatura conhecida em língua inglesa, em termos de Precision, Recall e F1, métricas utilizadas pelos autores em seus trabalhos.

Para realizar este trabalho, precisou-se portanto, de um dataset grande o suficiente e relevante que provesse além das informações necessárias, o padrão ouro. O melhor caminho para a obtênção de um conjunto de documentos com um padrão ouro de qualidade foi justamente criando um *dataset* de papers acadêmicos, pois estes normalmente já tem as palavras-chave que o autor julgou serem mais pertinentes e que melhor sintetizaram o artigo. Sendo assim, o padrão ouro era facilmente obtido lendo as palavras-chave do autor.

Foi criado um conjunto de 200 documentos a partir de 200 artigos acadêmicos, extraindo de cada um o título, o *abstract* e as palavras-chave.

Os algoritmos utilizados no contexto deste trabalho foram FirstPhrases, YAKE, TextRank, SingleRank, TopicRank, PositionRank e MultipartiteRank. Em seguida, bastava comparar as palavras-chave oriundas de cada algoritmo extrator e comparar com as palavras-chave do padrão ouro. Com isso, foi possível criar as métricas necessárias para as devidas comparações com a ampla literatura obtida em língua inglesa.

2 Trabalhos Relacionados

É possível encontrar uma vasta gama de trabalhos que abordam algoritmos geradores de palavras-chave, contudo, para o escopo deste trabalho, infelizmente, todos os trabalhos relacionados são em língua inglesa.

Alguns trabalhos utilizam algoritmos supervisionados para gerar as palavras-chave. Neste contexto é necessária um extenso conjunto de dados, onde as palavras são classificadas como palavra-chave caso positivo, e não palavra-chave, caso negativo. Para destacar alguns, Chengzhi propôs a utilização do modelo Conditional Random Fields (CRF) para a extração das palavras (Ian H. Witten *et al.*, 1999), que segundo os autores, se mostram mais eficiente que outros modelos de aprendizado de máquina como os modelos lineares e modelos de máquina de suporte de vetores (SVM).

Outro modelo supervisionado proposto por Ian, chamado de Automatic Keyphrase Extraction ou KEA (Ian H. Witten *et al.*, 1999), utiliza a Máquina de Naïve Bayes como algoritmo de aprendizado para treinar o seu conjunto de dados e extrair as palavras-chave.

No âmbito dos algoritmos não supervisionados, existem diversos trabalhos realizados em língua inglesa. Muitos utilizam o algoritmo Text Rank, que é uma melhoria proposta por Rada do já consolidado algoritmo Page Rank (Rada Mihalcea and Paul Tarau, 2004). O Text Rank é um algoritimo de ranking baseado em grafos que em sua ideia central retorna os termos mais recomendados, ou mais votados como vértices.

Outra melhoria foi a proposta por Suhan Pan, em seu trabalho *An Improved TextRank Keywords Extraction Algorithm* (Suhan Pan, Zhiqiang Li, and Juan Dai, 2019). Neste TextRank melhorado, os autores escolhem como input para o TextRank um vetor de termos que eles já consideraram boas candidatas, pois utilizam o algoritmo TF-IDF e o algoritmo de entropia média para calcular os pesos das palavras de maior relevância no Texto.

Outro algoritmo que vale destaque, é o algoritmo PositionRank proposto por Corina Florescu e Cornelia Caragea (Corina Florescu and Cornelia Caragea, 2017). É proposta uma importante modificação em relação a um PageRank tradicional. Em um PageRank as palavras são rankeadas em relação ao documento inteiro, enquanto no PositionRank cada palavra possui um peso que relaciona todas as posições de ocorrencia da palavra. Esta palavra sim, é incorporada em um PageRank enviesado, que atribui uma "preferência" diferente para cada palavra.

3 Modelos para extração automática de palavras-chave

No presente trabalho, foram comparados os modelos de aprendizado não-supervisionado estatístico: FirstPhrases e YAKE. E modelos de aprendizado não-supervisionado baseado em grafos: TextRank, SingleRank, TopicRank, PositionRank e MultipartiteRank.

Nesta seção, o funcionamento de cada modelo será melhor detalhado.

3.1 Modelos de aprendizado não-supervisionado estatístico

Os métodos estatísticos se utilizam de métricas como frequência de cada palavra e co-ocorrência para determinar as palavras mais relevantes em um documento. Alguns métodos utilizam características simbólicas mais avançadas, como TF-IDF e YAKE!.

O método FirstPhrases é usado como linha de base. Ele simplesmente extrai as primeiras palavras do texto, ignorando as stopwords.

O método YAKE utiliza como base certas características do texto para pontuar a relevância das palavras. São 5 características. Casing, considera que termos em letras maiúsculas tendem a ser mais relevantes. Posição do termo, termos que estão mais no começo do texto tendem a ser mais relevantes. Frequência dos termos, termos que aparecem mais vezes em um texto tendem a ser mais relevantes. Relação do termo com o contexto, mede com quantos termos diferentes o candidato co-ocorre, termos mais relevantes co-ocorrem com menos termos diferentes. Termo frase diferente, termos que aparecem em múltiplas frases tendem a ser mais relevantes (Ricardo Campos *et al.*, 2010)

3.2 Modelos de aprendizado não-supervisionado baseados em grafos

Métodos baseados em grafos geram um grafo de termos relacionados no documento. São usados métodos de rankeamento de grafos para computar a importância de cada nó – palavra – no grafo.

No método TextRank, a conexão entre os nós no grafo ocorre a partir da co-ocorrência de palavras dentro de uma janela (o padrão é w=2). O grafo é não direcionado e não ponderado. O algoritmo utilizado para determinar a importância de cada nó é o PageRank, do Google, é uma algoritmo recursivo, onde a importância de cada nó é determinada pela importância dos nós que se relaciona(Rada Mihalcea and Paul Tarau, 2004)).

O método SingleRank utiliza a mesma lógica e etapas que o TextRank, com a diferença de que o algoritmo PageRank é computado considerando os relacionamentos entre nós de forma ponderada e uma maior janela w = 10 (Xiaojun Wan and Jianguo Xiao, 2008).

TopicRank utiliza uma abordagem distinta, onde cada nó no grafo é um tópico. Um tópico é identificado extraindo frases nominais, contendo a maior sequência de substantivos e adjetivos. O relacionamento entre os diferentes tópicos é dado

pela sobreposição de palavras contidas em cada um, quanto maior a sobreposição, mais similares são os tópicos. Após o relacionamento entre os tópicos, é utilizado o mesmo algoritmo de ranqueamento do TextRank para rankear os tópicos. Para a seleção das palavras-chave, os tópicos mais importantes são selecionados e as palavras mais frequentes no tópico são selecionadas (Adrien Bougouin, Florian Boudin, and Beatrice Daille, 2013).

No PositionRank, o grafo é construído assim como no TextRank, com as palavras sendo nós e os relacionamentos sendo co-ocorrências entre as palavras em uma janela de 10 palavras (w = 10), o peso de cada aresta é dado pelo número de vezes que as palavras apareceram na mesma janela. A principal diferença para o TextRank é no cálculo do rankeamento usando um algoritmo Pagerank enviesado pela posição das palavras no texto. A principal ideia é que palavras no começo do texto tendem a ser mais relevantes(Corina Florescu and Cornelia Caragea, 2017).

MultipartiteRank é baseado no TopicRank, considerando frases-chave em tópicos como nós no grafo. O relacionamento é direcionado entre as frases-chave é dado se elas pertencerem a diferentes tópicos e o peso é computado de acordo com a distância entre elas no documento. O algoritmo de ranqueamento das frases-chave é o mesmo que o TextRank (Adrien Bougouin, Florian Boudin, and Beatrice Daille, 2013).

4 Experimentos e Resultados

4.1 Base de dados

Os experimentos foram realizados utilizando como base de dados 197 artigos científicos publicados no portal SciELO - Brazil [1]. Os artigos foram obtidos buscando-se pelos termos: educação à distância; educação na pandemia; inclusão digital; jogos educacionais; e mobile learning. Na Tabela 1 está a distribuição de artigos em cada um dos termos.

Foram utilizados os títulos e resumos concatenados como textos de entrada para os algoritmos extratores de palavras-chave. Além disso, as palavras-chave dos autores foram utilizadas como o alvo para comparar os resultados dos algoritmos.

Todo o processo de extração de informações foi feito de forma manual pelos autores deste artigo.

A base de dados encontra-se disponível publicamente, assim como todo o código utilizado nos experimentos [2].

Termo	#Documentos
Educação à distância	15
Educação na pandemia	83
Inclusão digital	29
Jogos educacionais	30
Mobile learning	40

Tabela 1: Distribuição de documentos por termo de busca.

4.2 Métricas de avaliação

Foram utilizadas as seguintes métricas para avaliação dos experimentos: Mean Reciprocal Rank (MRR), Precision, Recall e F1. Visto que são utilizadas em trabalhos prévios da literatura (Rada Mihalcea and Paul Tarau, 2004; Xiaojun Wan and Jianguo Xiao, 2008; Kazi Saidul Hasan and Vincent Ng., 2010).

Essas métricas foram computadas utilizando diferentes rankings (k) dos resultados dos modelos (desempenho@k). Os valores de k utilizados variaram de 1 a 10. Por exemplo: precision@4 considera apenas as 4 primeiras palavras-chave retornadas pelo modelo para calcular a precisão.

A referência de alvo para comparação das palavras-chave relevantes são as palavras-chave que os autores de cada documento determinaram.

4.3 Resultados e Discussão

O objetivo principal dos experimentos foi determinar quais modelos de extração de palavras-chave melhor desempenham em textos científicos de língua portuguesa.

Para isso, foram comparados os modelos de aprendizado não-supervisionado estatístico: FirstPhrases e YAKE. E modelos de aprendizado não-supervisionado baseado em grafos: TextRank, SingleRank, TopicRank, PositionRank e MultipartiteRank.

Todos os modelos utilizaram as mesmas etapas de processamento de dados: remoção de palavras

[1] https://www.scielo.br

[2] https://github.com/CostaFernando/keyword-extraction-portuguese

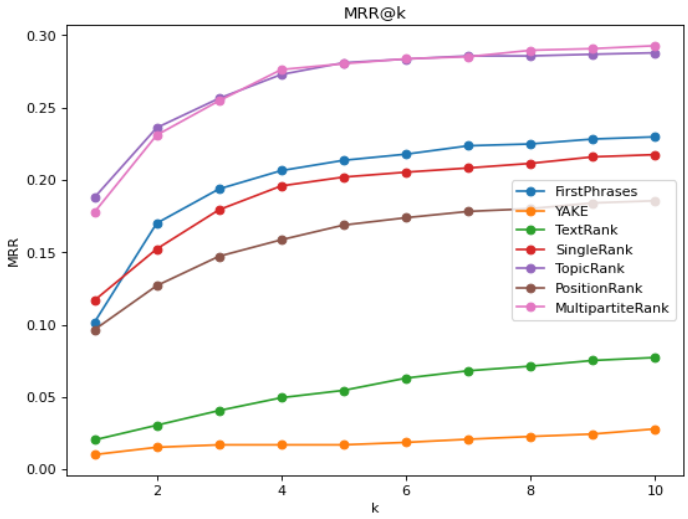

Figura 2: Distribuição de documentos por termo de busca.

comuns do Português (stopwords), pontuação e sinais de parênteses.

Além disso, as configurações utilizadas em cada modelo foram as configurações padrão nas implementações dos modelos da biblioteca pke - python keyphrase extraction[3].

Na figura 2, são comparados diferentes modelos da literatura utilizando a métrica MRR em diferentes rankings. Os modelos que obtêm melhores desempenhos são TopickRank e MultipartiteRank, chegando a 0,293 com o MultipartiteRank para k = 10.

Os modelos YAKE e TextRank obtiveram os resultados mais baixos na métrica MRR, com o YAKE chegando a um valor máximo de MRR de 0,028 para k = 10.

Além disso, foram medidos precision (P), recall (R) e F1-score (F1) com k variando de 1 a 10 para todos os modelos.

O modelo MultipartiteRank obteve de forma geral os melhores valores para as métricas P, R e F1. Atingindo F1 máximo de 13,60% para k = 4. Como pode ser observado na figura 3 e tabela 2.

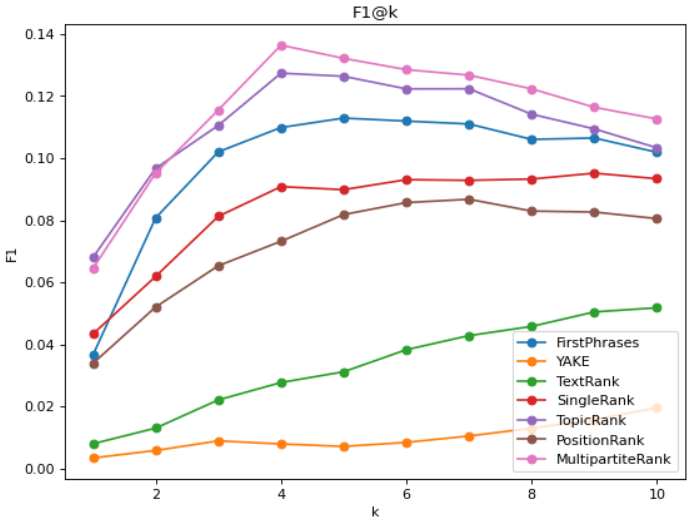

Figura 3: Curvas F1 para diferentes modelos.

[3] https://github.com/boudinfl/pke

Modelo	Top2			Top4			Top6			Top8		
	p_2	r_2	f1_2	p_4	r_4	f1_4	p_6	r_6	f1_6	p_8	r_8	f1_8
FirstPhrases	12,90%	5,90%	8,10%	11,80%	10,50%	11,00%	9,90%	13,20%	11,20%	8,40%	14,80%	10,60%
YAKE	1,00%	0,40%	0,60%	0,90%	0,70%	0,80%	0,80%	0,90%	0,80%	1,00%	1,80%	1,30%
TextRank	2,00%	1,00%	1,30%	2,90%	2,70%	2,80%	3,30%	4,70%	3,80%	3,60%	6,60%	4,60%
SingleRank	9,90%	4,60%	6,20%	9,60%	8,80%	9,10%	8,10%	11,20%	9,30%	7,30%	13,30%	9,30%
TopicRank	**16,00%**	**7,00%**	**9,70%**	13,70%	12,10%	12,70%	10,80%	14,40%	12,20%	9,00%	16,00%	11,40%
PositionRank	8,60%	3,80%	5,20%	7,70%	7,10%	7,30%	7,40%	10,30%	8,60%	6,50%	11,80%	8,30%
MultipartiteRank	15,70%	6,90%	9,50%	**14,70%**	**13,00%**	**13,60%**	**11,30%**	**15,20%**	**12,80%**	**9,60%**	**17,20%**	**12,20%**

Tabela 2: Comparativo de diferentes modelos em termos de Precision (P), Recall (R) e F1-score para k = 2, 4, 6, 8. Os melhores resultados estão destacados em azul.

YAKE obteve o desempenho mais baixo, com 1,3% F1 para k = 8.

4.4 Considerações sobre os desempenhos dos modelos

Ao analisar o desempenho dos modelos, deve-se levar em consideração os dados utilizados como entrada. Foram resumos de artigos científicos, que normalmente contém maior densidade de informação, por serem mais curtos e conterem os termos mais relevantes.

O desempenho desses modelos pode ser diferente quando aplicados em textos completos, com maior quantidade de palavras e menos densidade de informação.

Isso pode ser verificado com o desempenho relativamente alto alcançado pelo algoritmo FirstPhrases. Esse era para ser um método apenas usado como linha de base para as métricas, dada sua simplicidade. No entanto, ele conseguiu se sair melhor do que outros métodos, acertando palavras-chave relevantes citadas pelos autores dos artigos. Isso provavelmente se deve ao fato de os resumos dos artigos conterem maior densidade de palavras-chave do que o texto completo, favorecendo esse tipo de método.

Em contraste, o método YAKE obteve os mais baixos resultados. Indicando uma possível dificuldade de funcionar com textos menores, com maior densidade de palavras-chave, como os resumos dos artigos. Os resultados contrastam com os obtidos na literatura (Campos et al., 2020).

5 Conclusão e Trabalhos Futuros

Este artigo teve o objetivo de propor uma abordagem ainda muito inexplorada, que é a de trabalhos que abordem a extração de palavras-chave de textos em língua portuguesa. Um grande desafio aqui foi de obter uma boa base devidamente classificada, fidedígnea e com uma boa base científica.

Com os 200 artigos em língua portuguesa, devidamente classificados com as palavras-chaves fornecidas pelos autores, foi possível fazer todos os experimentos utilizando os algoritmos já conhecidos como bons extratores de palavras-chave. E pode ser comprovada a eficácia destes classificadores utilizando todos os critérios e métricas supracitados.

Tentamos utilizar tais técnicas para gerar questões do tipo “complete as lacunas com a expressão adequada”, na qual utilizamos os extratores para ver 10 termos-chaves mais importantes em um texto, e gerar uma questão de prova utilizando a frase e o contexto em que se encontra esta palavra-chave. Contudo, os resultados não foram ótimos, e não existe uma boa literatura para comparação de resultados ainda. Como trabalho futuro gostaríamos de conseguir uma boa base classificada de questões desse tipo, de preferência de um banco de provas, para tentar alterar algum dos algoritmos não supervisionados abordados neste trabalho, ou mesmo partir para uma abordagem supervisionada.

References

Adrien Bougouin, Florian Boudin, and Beatrice Daille (2013) 'TopicRank: Graph-Based Topic Ranking for Keyphrase Extraction', in. *Proceedings of the Sixth International Joint Conference on Natural Language Processing*.

Corina Florescu and Cornelia Caragea (2017) 'PositionRank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents', in. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Ian H. Witten *et al.* (1999) 'KEA: Practical Automatic Keyphrase Extraction', in. *Conference: Proceedings of the Fourth ACM conference on Digital Libraries, August 11-14, 1999, Berkeley, CA, USA*.

Kazi Saidul Hasan and Vincent Ng. (2010) 'Conundrums in unsupervised keyphrase extraction: making sense of the state-of-the-art.', in *Conundrums in unsupervised keyphrase extraction: making sense of the state-of-the-art. In Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 365–373.

Rada Mihalcea and Paul Tarau (2004) 'TextRank: Bringing Order into Texts', in. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*.

Ricardo Campos *et al.* (2010) 'YAKE! Keyword extraction from single documents using multiple local features', in. *Information Sciences: an International JournalVolume 509Issue C*.

Romilda Guilland *et al.* (2021) 'Prevalência de sintomas de depressão e ansiedade em trabalhadores durante a pandemia da Covid-19', in. *Epidemiol. Serv. Saúde vol.29 no.4 Brasília set. 2020 Epub 20-Ago-2020*.

Suhan Pan, Zhiqiang Li, and Juan Dai (2019) 'An Improved TextRank Keywords Extraction Algorithm', in. *ACM TURC '19: Proceedings of the ACM Turing Celebration Conference - China*.

Xiaojun Wan and Jianguo Xiao (2008) 'CollabRank: Towards a Collaborative Approach to Single-Document Keyphrase Extraction', in. *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*.