4. Experimento

Para avaliar o **GRNews** e testar a acurácia do processo de recomendação de matérias relacionadas, desenvolvemos um conjunto de experimentos que certificam os módulos do sistema sobre um corpus de matérias já relacionadas manualmente.

Primeiro, detalharemos como esse corpus de matérias foi obtido, depois definiremos uma heurística para avaliação e, em seguida, executaremos uma função especialmente desenvolvida para combinar as diversas *features* obtidas no Capítulo 3 de modo a descobrir qual combinação possui a maior assertividade sobre o corpus.

Ao final, apresentaremos os resultados obtidos e discutiremos as possibilidade de melhoria sobre o sistema **GRNews**.

4.1. O corpus

O corpus utilizado pertence ao portal de noticias **G1** da **Globo.com** e possui um total de 492 mil matérias, produzidas no período de fevereiro de 2010 a outubro de 2011, divididas em 162 editorias.

Deste corpus, 72 mil matérias possuem o componente de matéria relacionada, o "**saibamais**". Este componente mantém os links das matérias que foram relacionadas pelo editor.

Para diminuir o tempo de execução dos algoritmos, reduzimos o corpus para um total de 14.400 matérias já relacionadas, o que representa 20% do total de matérias relacionadas manualmente. Também reduzimos as representações das editorias, ficando apenas com as 15 principais editorias, de acordo com o seu número de matérias publicadas.

Decidimos distribuir de forma igualitária as matérias em suas editorias, de modo a promover uma visão das editorias com maior percentual de acertos, conforme mostra a Tabela 10.

Tabela 10 – Representação do corpus

1	Brasil	960
2	São Paulo	960
3	Rio de Janeiro	960
4	Minas Gerais	960
5	Economia	960
6	Política	960
7	Mundo	960
8	Espírito Santo	960
9	Pop & Arte	960
10	Auto Esporte	960
11	Concursos e Emprego	960
12	Ciência e Saúde	960
13	Música	960
14	Mercados	960
15	Tecnologia e Games	960
	TOTAL	14.440

Foram necessárias etapas de limpeza do corpus para garantir que todos os links de matérias apontavam para matérias existentes na base de dados. Isso se fez necessário porque o cadastro de matérias relacionadas é aberto e permite a ligação a matérias externas. Assim sendo, matérias que possuíam pelo menos um link externo foram descartadas durante a preparação do corpus.

Ao final desta etapa de limpeza, temos então um corpus de matérias relacionadas manualmente com links válidos e que corresponde a 20% do corpus total de matérias relacionadas manualmente.

Durante a etapa de limpeza, as matérias foram indexadas no servidor de busca **SOLR** que dá suporte ao sistema **GRNews**.

4.2. Critério para avaliação do sistema

Para avaliar o sistema foi necessária a construção de um módulo de combinação das *features*, extraídas pelo extrator de *features,* como discutido no Capítulo 3.

Para cada combinação de *features***,** o módulo percorre o corpus com matérias previamente relacionadas e realiza, para cada matéria, uma recomendação de *n* matérias, onde *n* é o número de matérias que serão

recomendadas. Após a recomendação, o sistema compara as matérias que foram relacionadas manualmente com as que foram recomendadas pelo **GRNews** e, em seguida, extrai a interseção. Se a interseção não for nula, consideramos que a recomendação foi válida. Esta métrica é conhecido em information retrieval como **precision**. Em nosso trabalho utilizamos P@5(precisão com 5 recomendações) e or P@10 (precisão com 10 recomendações) para realizar a verificação da nossa abordagem.

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

Figure 11 – Fórmula de Precisão

4.3. Resultados obtidos

Para avaliar a hipótese de melhoria da recomendação com *features* extraídas do texto, estipulamos um *baseline* para o projeto **GRNews**. Considerando apenas os *unigrams* mais relevantes do texto, submetemos o sistema **GRNews** ao processamento de um conjunto de recomendações contra o corpus de teste. Com o resultado das recomendações, conseguimos estabelecer um percentual de aproximadamente 38% de acurácia, como pode ser observado na Tabela 11. A partir deste *baseline*, iniciamos um conjunto de tentativas onde variamos as *features*, o número de matérias recomendadas e a utilização da medida de similaridade para tentar aumentar a acurácia do **GRNews** nas recomendações. A seguir vamos apresentar os resultados obtidos com cada rodada de recomendações a fim de estudar os dados e eleger as melhores combinações de *features* para o **GRNews**.

A primeira tentativa foi descobrir a melhor combinação de critérios sobre todo o corpus de teste, utilizando como base um total de 5 recomendações e ainda dispensando o fator de similaridade das matérias, uma vez que ele poderia não ser determinante para encontrarmos a melhor relação. Com isso obtivemos os resultados mostrados na Tabela 11.

Tabela 11 – Acurácia por *feature* com 5 recomendações sem similaridade

Total Geral					
ubtce	**0.471160528**	uthc	0.443363447	hce	0.408617095
ubthce	0.470465601	ubc	0.44266852	ce	0.405837387
uthce	0.46838082	ubhc	0.44266852	bt	0.40514246
utce	0.466990966	tce	0.441973593	bth	0.403752606
ubte	0.46143155	thce	0.441278666	he	0.397498263
ubthe	0.46143155	bte	0.439888812	e	0.395413482
ubtc	0.460736623	bthe	0.439193885	uhc	0.394718555
ubthc	0.460736623	ut	0.435719249	uc	0.389854065
ubhce	0.457956915	uth	0.434329395	uh	0.389159138
ubce	0.45656706	bhce	0.432244614	**u**	**0.384294649**
uthe	0.45656706	uhe	0.432244614	thc	0.382209868
ute	0.455872133	ub	0.431549687	tc	0.380125087
ubt	0.451702571	ue	0.431549687	bc	0.375955525
ubth	0.451702571	ubh	0.43085476	bhc	0.375955525
btce	0.448922863	bce	0.429464906	bh	0.357887422
ubhe	0.448922863	te	0.427380125	th	0.357887422
bthce	0.448227936	the	0.426685198	t	0.357192495
ube	0.446838082	bhe	0.422515636	b	0.355107714
uhce	0.446838082	be	0.419735928	hc	0.161223072
uce	0.445448228	btc	0.419735928	c	0.136900625
utc	0.444753301	bthc	0.419735928	h	0.03752606

Nota: u – unigrams, b – bigrams, t – trigrams, h – html tags, c – captions, e – entidades.

Para entender as combinações na Tabela 11 e na demais, é necessária a interpretação dos rótulos, utilizando a legenda mostrada no rodapé da Tabela 11.

Neste primeiro resultado, observamos que, utilizando as *features* com a combinação **ubtce** (ou seja, sem HTML tags), conseguimos uma acurácia de aproximadamente 47%, o que representa uma melhoria de aproximadamente 24% sobre o *baseline*.

Se olharmos a distribuição por editoria detalhadas no Apêndice A, observamos que *São Paulo* e *Rio de Janeiro* obtiveram as melhores taxas de assertividade, com aproximadamente 78% e 82%, respectivamente. Levando-se em consideração a combinação mínima para estas categorias, temos, para *São Paulo* a sequência **ubt** enquanto que, para o *Rio de Janeiro*, temos a sequência **ubce**.

Da mesma forma, temos também as editorias com o pior percentual de acurácia, que são: *Mercados*, com aproximadamente 19%; *Economia*, com 28%; e *Minas Gerais*, com aproximadamente 27%. Nestas categorias as melhores sequências de features foram: **ubt**, **utce** e **ube**.

Uma primeira conclusão possível sobre estas informações está no fato de que, em editorias mais organizadas, ou seja, com menos conteúdo diversificado, a recomendação é mais assertiva. O contexto de *Economia* é mais abrangente que os contextos de *Rio de Janeiro* e *São Paulo*. *Economia* apresenta matérias do Brasil e do mundo, enquanto as editorias *Rio de Janeiro* e *São Paulo* são focadas no noticiário local, o que torna o contexto muito próximo.

Uma outra hipótese é a qualidade da equipe de jornalistas por editoria. Como *Rio de Janeiro* e *São Paulo* têm mais visibilidade, os jornalistas dedicam uma maior atenção ao relacionamento do conteúdo. Esta hipótese explicaria o fato de *Economia*, *Mercados e Minas Gerais* apresentarem uma acurácia baixa.

A segunda tentativa foi em razão da dúvida sobre o fator de similaridade. Durante este teste apenas habilitamos o fator de similaridade para ajustar o *score* das matérias com maior similaridade textual. Com isso, pudemos observar quanto o fator de similaridade entre as matérias influencia o nosso algoritmo. O novo *score* é dado pela multiplicação simples entre o *score* anterior e o fator de similaridade entre as duas matérias, dado pela distância dos cossenos.

Tabela 12 – Acurácia por *feature* com 5 recomendações com similaridade

Total Geral							
ubthce	**0.49409312**		bthce	0.478109798		u	0.457261987
ubtce	0.493398193		uhce	0.476719944		uh	0.457261987
utce	0.492008339		uthc	0.476025017		btc	0.451007644
uthce	0.492008339		bthe	0.47533009		bthc	0.451007644
ubce	0.491313412		uce	0.47533009		bth	0.44266852
ubhce	0.491313412		utc	0.47533009		hce	0.44266852
ubthe	0.489228631		bte	0.474635163		bt	0.441973593
ubte	0.488533704		thce	0.473245309		ce	0.441278666
ubhe	0.48714385		bhce	0.472550382		he	0.438498958
ubtc	0.48714385		tce	0.471855455		e	0.436414177
ube	0.486448923		bce	0.471160528		tc	0.426685198
ubc	0.485753996		the	0.471160528		thc	0.424600417
ubthc	0.485059069		te	0.470465601		bhc	0.41695622
ubhc	0.484364142		ut	0.469770674		t	0.41695622
ute	0.484364142		uth	0.469075747		th	0.415566366
uthe	0.484364142		bhe	0.467685893		bc	0.414871438
ub	0.479499653		uhe	0.467685893		bh	0.403057679
ubt	0.479499653		be	0.466990966		b	0.400972898
ubh	0.478804726		uc	0.466296039		hc	0.134120917
ubth	0.478804726		ue	0.466296039		c	0.113273106
btce	0.478109798		uhc	0.465601112		h	0.031966644

Como podemos observar na Tabela 12, o uso do fator de similaridade entre os textos das matérias implica em uma melhoria de aproximadamente 29% sobre o *baseline* estabelecido.

Olhando para a Tabela 12, destacamos a mudança na sequência de *features* combinadas que obteve a melhor acurácia. Saímos da sequência **ubtce** para a sequência **ubthce.** Com o fator de similaridade, as informações extraídas de tags HTML informativas ganharam mais relevância.

Observando a distribuição por editoria, podemos destacar uma melhora no percentual em quase todas as editorias, com exceção de: *Brasil*, que teve uma ligeira queda, e *Ciência e Saúde*, que não variou com o fator de similaridade aplicado. Ainda, observando as editorias, assim como na Tabela 11, quase todas as combinações vencedoras foram mudadas em relação ao teste executado sem o fator de similaridade.

Até este ponto, estávamos realizando apenas 5 recomendações por matéria.

Nas tabelas a seguir, verificaremos qual o percentual de melhora que será obtido se aumentarmos o número de itens recomendados de 5 para 10. Executaremos então os dois procedimentos anteriores, porém alterando o número de recomendações para 10.

Tabela 13 – Acurácia por feature com 10 recomendações sem similaridade

Total Geral							
uthce	**0.53856845**		uthc	0.520500347		bth	0.483669215
ubthce	0.537873523		uce	0.51980542		uhc	0.483669215
ubtce	0.537178596		utc	0.51980542		bt	0.482974288
utce	0.537178596		btce	0.517720639		uc	0.481584434
ubthc	0.534398888		uhe	0.516330785		uh	0.47533009
ubc	0.533703961		bthce	0.515635858		hce	0.474635163
ubtc	0.533703961		ue	0.514246004		ce	0.473940236
ubhc	0.533009034		ut	0.51285615		**u**	**0.470465601**
ubhce	0.532314107		uth	0.51285615		he	0.469770674
ubt	0.532314107		bte	0.510771369		e	0.469075747
ubth	0.532314107		thce	0.509381515		bc	0.447533009
ubthe	0.532314107		bthe	0.508686588		bhc	0.446143155
ubte	0.53161918		tce	0.508686588		thc	0.444058374
ubh	0.530924253		bhce	0.499652536		tc	0.443363447
ub	0.529534399		the	0.499652536		bh	0.428075052
ubce	0.529534399		te	0.498957609		b	0.426685198
uthe	0.529534399		bce	0.498262682		th	0.425295344
ute	0.528839472		bhe	0.493398193		t	0.42390549
ubhe	0.526059764		btc	0.492703266		hc	0.143849896
ube	0.523280056		bthc	0.492008339		c	0.122307158
uhce	0.521890202		be	0.491313412		h	0.029186935

Com 10 recomendações, a acurácia alcança aproximadamente 54%. Porém, podemos perceber que o *baseline* para 10 recomendações também é alto, sendo cerca de 47%. Desta forma, é possível notar que o aumento percentual com o uso de outras *features* é de 15%, o que é um ganho menor quando comparado com o mesmo teste com 5 recomendações. Isto significa que, com 10 recomendações, é mais fácil acertar, até mesmo fazendo uso de *features* simples, como é o caso do *baseline*.

Mas, de qualquer forma para efeito de recomendação, é correto dizer que

conseguimos um grau maior de precisão em matérias relacionadas quando aumentamos o número de matérias recomendas. Quando observamos, por exemplo, a recomendação por editoria, percebemos que a editoria *Rio de Janeiro* alcança uma acurácia de aproximadamente 91%.

O experimento seguinte apenas comprova o que já foi visto anteriormente na recomendação com utilização do fator de similaridade. Como pode ser visto na Tabela 14, o ganho percentual sobre o *baseline* aumenta para aproximadamente 21%.

Tabela 14 – Acurácia por *feature* com 10 recomendações com similaridade

Total Geral							
ubtce	**0.566365532**	ubh	0.548992356	bthc	0.520500347
ubthce	0.565670605	uhce	0.548992356	btc	0.519110493
Ubte	0.564975678	ub	0.546907575	u	0.515635858
ubthe	0.564280751	utc	0.546212648	uh	0.513551077
uthce	0.562195969	uthc	0.544127867	bth	0.512161223
Utce	0.561501042	bte	0.54343294	bt	0.510076442
ubce	0.560806115	bthe	0.542738013	hce	0.501737318
ubthc	0.560806115	ue	0.542738013	ce	0.500347464
Ube	0.560111188	uhe	0.541348158	he	0.500347464
ubtc	0.559416261	bhce	0.53856845	e	0.498262682
ubhce	0.558721334	bce	0.537873523	tc	0.489923558
ubhe	0.558026407	thce	0.537873523	thc	0.487838777
ute	0.558026407	tce	0.536483669	bhc	0.472550382
uthe	0.558026407	the	0.533703961	t	0.472550382
ubth	0.556636553	be	0.533009034	th	0.471855455
ubhc	0.554551772	bhe	0.533009034	bc	0.469770674
ubt	0.554551772	te	0.533009034	bh	0.460736623
ubc	0.553161918	ut	0.533009034	b	0.459346769
uce	0.55038221	uth	0.53161918	hc	0.157053509
btce	0.549687283	uc	0.526059764	c	0.131341209
bthce	0.548992356	uhc	0.523974983	h	0.035441279