

UNIVERSIDADE FEDERAL DO AMAZONAS
INSTITUTO DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

JANIEL MEDEIROS DOS SANTOS

UMA ESTRATÉGIA DE RECOMENDAÇÃO ASSOCIATIVA DE
ETIQUETAS USANDO GRAFO DE CONTEXTO EM ESTADO DE
COLD START

Manaus-AM

2017

JANIEL MEDEIROS DOS SANTOS

UMA ESTRATÉGIA DE RECOMENDAÇÃO ASSOCIATIVA DE
ETIQUETAS USANDO GRAFO DE CONTEXTO EM ESTADO DE
COLD START

Dissertação apresentada ao Programa de Pós-graduação em Informática da Universidade Federal do Amazonas como requisito parcial para obtenção do grau de Mestre em Informática.

Área de concentração: Redes de computadores e telecomunicações.

Orientador:

Prof. Dr. César Augusto Viana Melo

Manaus-AM

2017

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

S237u Santos, Janiel Medeiros dos
Uma estratégia de recomendação associativa de etiquetas
usando grafo de contexto em estado de cold start / Janiel Medeiros
dos Santos. 2017
87 f.: il. color; 31 cm.

Orientador: César Augusto Viana Melo
Dissertação (Mestrado em Informática) - Universidade Federal do
Amazonas.

1. Etiquetagem. 2. Recomendação de etiquetas. 3. Método
associativo. 4. Grafo de contexto. 5. Cold start. I. Melo, César
Augusto Viana II. Universidade Federal do Amazonas III. Título



PODER EXECUTIVO
MINISTÉRIO DA EDUCAÇÃO
INSTITUTO DE COMPUTAÇÃO



UFAM

PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

FOLHA DE APROVAÇÃO

"Uma estratégia de recomendação associativa de etiquetas usando grafo de contexto em estado de cold start"

JANIEL MEDEIROS DOS SANTOS

Dissertação de Mestrado defendida e aprovada pela banca examinadora constituída pelos Professores:

Prof. César Augusto Viana Melo - PRESIDENTE

Prof. Edlêno Silva de Moura - MEMBRO INTERNO

Prof. Moisés Gomes de Carvalho - MEMBRO EXTERNO

Manaus, 02 de Outubro de 2017

Agradecimentos

Meus sinceros agradecimentos:

- a Deus, pelo ânimo, discernimento e perseverança dados durante a caminhada;
- ao Prof. César Melo, pela motivação, atenção e disponibilidade nas orientações;
- a todos os professores e demais funcionários do ICOMP/UFAM, pelo esforço e dedicação na formação e inovação científica da região;
- à Fundação de Amparo à Pesquisa do Estado do Amazonas e a Universidade Federal do Amazonas, pelo suporte financeiro, sem esse incentivo seria impossível me manter;
- a minha mãe, Maria Irismar, por ter investido em minha educação, e por ter me dado força e chamado a atenção quando era necessário, agradeço por tudo;
- a minha namorada Gabriela pela compreensão, quando não pude estar presente em alguns momentos;
- a todos os meus colegas, amigos e pessoas que contribuíram direta e indiretamente para que eu pudesse finalizar essa etapa.

*“A vitória mais bela
que se pode alcançar
é vencer a si mesmo.”*

Santo Inácio de Loyola

Resumo

A crescente produção de conteúdo multimídia na web tem estimulado estudos relacionados ao aprimoramento das estratégias responsáveis pela organização e recuperação desse conteúdo nas aplicações. Dentre as diversas estratégias existentes, a atribuição de um conjunto de etiquetas a um objeto expressando o seu conteúdo, denominada de processo de etiquetagem, têm sido objeto de estudos recentes, principalmente os métodos associativos que exploram padrões de coocorrência. Apesar dos métodos associativos apresentarem resultados estado da arte, poucos exploram o comportamento desse tipo de recomendação em um estado em que o objeto está sendo inserido no sistema e não possui informações previamente associadas, caracterizado como problema de *cold start*. Investigar o comportamento das estratégias, em um estado de *cold start*, tem como motivação principal melhorar a qualidade das recomendações, principalmente quando as estratégias ficam dependentes das informações fornecidas pelos publicadores do conteúdo. Nesta dissertação apresenta-se uma estratégia de recomendação associativa de etiquetas que explora o conceito de grafos de contexto, por meio da integração da coocorrência e métricas de relevância de etiquetas com o conhecimento colaborativo dos relacionamentos existentes entre conceitos presentes nos artigos da Wikipédia. Com o equilíbrio na valorização das etiquetas candidatas a recomendação, a estratégia proposta obteve melhoria na qualidade da recomendação das etiquetas em vídeos que possuem uma quantidade mediana de etiquetas.

Palavras-chave: Etiquetagem, recomendação de etiquetas, método associativo, grafos de contexto e *cold start*.

Abstract

The growing production of multimedia content on the web has stimulated studies related to the improvement of the strategies responsible for the organization and recovery of this content in the applications. Among the several strategies, the assignment of a set of tags to an object expressing its content, called the tagging process, has been the subject of recent studies, mainly associative methods that are based on the exploration of co-occurrence patterns of tags. Although associative methods present state of the art results, few investigate the behavior of this type of recommendation in a state in which the object is being inserted into the system and has no previously associated information, characterized as cold start problem. To investigate the behavior of strategies in a cold start, its main motivation is to improve the quality of the recommendations, especially when the strategies are dependent on the information provided by the publishers of the content. This work presents an associative tags recommendation strategy that explores the concept of context graphs through the integration of co-occurrence and tags relevance metrics with the collaborative knowledge of existing relationships between concepts found in Wikipedia articles. With the balance in the valorization of the candidate tags for the recommendation, the proposed strategy obtained improvement in the quality of the recommendation of the tags in videos that have a median number of tags.

Keywords: Tagging, tag recommendation, associative methods, contexts graphs and cold start.

Lista de Figuras

1.1	Tráfego da rede na América Latina em horário de pico	p. 15
1.2	As 10 aplicações que mais geram tráfego no horário de pico na América Latina	p. 16
3.1	Cenário homogêneo.	p. 34
3.2	Cenário heterogêneo.	p. 35
3.3	Ambiente de experimentação automática.	p. 36
4.1	<i>Full Local Context Graph - flcg</i>	p. 46
4.2	<i>Degree Local Context Graph - dlcg</i>	p. 47
4.3	<i>Connected Local Context Graph - clcg</i>	p. 48
4.4	<i>Strong Local Context Graph - slcg</i>	p. 49
4.5	Expanded Full Local Contexts Graphs - <i>xflcg</i>	p. 50
4.6	Esquema da estratégia Sum^+wTS^+GC	p. 52
5.1	Ambiente de experimentação.	p. 73

Lista de Tabelas

2.1	Comparação das características dos métodos de recomendação de etiquetas.	p. 27
3.1	Parâmetros de análise das estratégias de recomendação.	p. 38
3.2	Desempenho do método Sum^+wTS com a base Youtube_ufmg.	p. 39
3.3	Desempenho do método Sum^+wTS com a amostra Youtube_Sports	p. 40
3.4	Desempenho do método Sum^+wTS com amostra Youtube_Sports_3000.	p. 41
3.5	Desempenho do método Sum^+wTS com a amostra Youtube_Sports_Rev.	p. 41
4.1	Desempenho do método Sum^+wTS^+GC com a amostra Youtube_Sports no estado sem <i>cold start</i>	p. 55
4.2	Desempenho do Sum^+wTS^+GC com a amostra Youtube_Sports no estado de <i>cold start</i>	p. 55
4.3	Impacto no desempenho da estratégia Sum^+wTS^+GC pelo uso dos grafos abrangentes.	p. 57
4.4	Impacto no desempenho da estratégia Sum^+wTS^+GC pelo uso dos grafos limitadores.	p. 58
4.5	Desempenho do método Sum^+wTS^+GC com a amostra Youtube_Sports_3000 em estado sem <i>cold start</i>	p. 58
4.6	Desempenho do método Sum^+wTS^+GC com a amostra Youtube_Sports_Rev em estado sem <i>cold start</i>	p. 60
4.7	Impacto no desempenho da estratégia Sum^+wTS^+GC pelo uso dos grafos abrangentes.	p. 62

4.8	Impacto no desempenho da estratégia Sum^+wTS^+GC pelo uso dos grafos limitadores.	p. 62
4.9	Impacto no desempenho da estratégia Sum^+wTS^+GC pelo uso dos grafos abrangentes.	p. 63
4.10	Impacto no desempenho da estratégia Sum^+wTS^+GC pelo uso dos grafos limitadores.	p. 63
4.11	Desempenho do método Sum^+wTS^+GC com a amostra Youtube_Sports_3000 em estado <i>cold start</i>	p. 63
4.12	Desempenho do método Sum^+wTS^+GC com a amostra Youtube_Sports_Rev em estado de <i>cold start</i>	p. 66
4.13	Impacto no desempenho da estratégia Sum^+wTS^+GC pelo uso dos grafos expandidos.	p. 68
4.14	Impacto no desempenho da estratégia Sum^+wTS^+GC pelo uso dos grafos limitadores.	p. 69
5.1	Perfil dos avaliadores	p. 73
5.2	Desempenho na avaliação qualitativa - Subconjuntos de etiquetas <i>menor</i> e <i>médio</i>	p. 75
5.3	Taxas de etiquetas escolhidas (taxa 1) e de etiquetas escolhidas que estão entre as cinco primeiras recomendadas (taxa 2).	p. 76
5.4	Desempenho do Sum^+wTS^+GC nas avaliações de acordo com o perfil do avaliador:	p. 76
5.5	Desempenho do Sum^+wTS^+GC nas avaliações de acordo com o perfil do avaliador (em cada grafo) :	p. 77
5.6	Taxa de etiquetas sugeridas que possuem artigo na Wikipédia:	p. 77
5.7	Taxa de etiquetas seleccionadas que possuem artigo Wikipédia:	p. 78
5.8	Taxa de recomendações que possuem etiquetas formais:	p. 78
5.9	Taxa de etiquetas formais que foram sugeridas (compõem o conjunto de recomendação):	p. 79
5.10	Taxa de avaliações que tinham etiquetas formais:	p. 79

5.11	Taxa de etiquetas formais que foram escolhidas (compõem o conjunto de etiquetas escolhidas):	p. 79
5.12	Taxa de etiquetas formais sugeridas que estão entre as cinco primeiras recomendadas (conjunto de recomendação):	p. 80
5.13	Taxa de etiquetas formais sugeridas que estão entre as seis primeiras recomendadas (conjunto de recomendação):	p. 80
5.14	Taxa de etiquetas formais sugeridas que estão entre as sete primeiras recomendadas (conjunto de recomendação):	p. 81
5.15	Taxa de etiquetas formais escolhidas que estão entre as cinco primeiras sugeridas:	p. 81
5.16	Taxa de etiquetas formais escolhidas que estão entre as seis primeiras sugeridas:	p. 81
5.17	Taxa de etiquetas formais escolhidas que estão entre as sete primeiras sugeridas:	p. 82

Sumário

1	Introdução	p. 13
1.1	Problema	p. 14
1.2	Motivação	p. 15
1.3	Objetivo	p. 16
1.4	Principais Contribuições	p. 17
1.5	Estrutura da Dissertação	p. 17
2	Conceitos e Trabalhos Relacionados	p. 19
2.1	Conceitos	p. 20
2.1.1	Recomendação de etiquetas	p. 20
2.1.2	Método associativo	p. 21
2.1.3	<i>Cold start</i>	p. 21
2.2	Métodos associativos	p. 22
2.3	Métodos associativos e o <i>cold start</i>	p. 23
2.4	Outros métodos de recomendação de etiquetas e o <i>cold start</i>	p. 25
2.5	Considerações Finais do Capítulo 2	p. 28
3	Método associativo de recomendação de etiquetas Sum^{+wTS}	p. 29
3.1	Sum^{+wTS}	p. 30

3.2	As Bases de dados	p. 32
3.3	Metodologia de Avaliação	p. 33
3.4	Ambiente de Experimentação	p. 36
3.5	Resultados numéricos	p. 37
3.5.1	Primeira rodada - amostra Youtube_ufmg	p. 38
3.5.2	Segunda rodada - amostra Youtube_Sports	p. 39
3.5.3	Terceira rodada - amostra Youtube_Sports revisada	p. 40
3.6	Considerações Finais do Capítulo 3	p. 41
4	Estratégia de recomendação associativa de etiquetas usando grafos de contexto	p. 44
4.1	Encontrando conceitos em textos curtos	p. 45
4.1.1	Full Local Context Graph - <i>flcg</i>	p. 45
4.1.2	Degree Local Context Graph - <i>dleg</i>	p. 47
4.1.3	Connected Local Context Graph - <i>clcg</i>	p. 47
4.1.4	Strong Local Context Graph - <i>slcg</i>	p. 48
4.1.5	Expanded Contexts Graphs	p. 49
4.1.6	Algoritmos de análise de <i>links</i>	p. 50
4.2	Estratégia de recomendação associativa de etiquetas com grafos de contexto	p. 51
4.2.1	Extração dos termos	p. 52
4.2.2	Método associativo	p. 52
4.2.3	Grafo de contexto	p. 53
4.3	Resultados Numéricos	p. 54
4.3.1	Primeira rodada - amostra Youtube_Sports	p. 54
4.3.2	Segunda Rodada - amostra Youtube_Sports revisada	p. 55
4.4	Discussões Sobre o Capítulo 4	p. 69
5	Avaliação qualitativa da estratégia de recomendação associativa de	

etiquetas usando grafos de contexto	p. 71
5.1 Ambiente de experimentação	p. 72
5.2 Resultados numéricos	p. 73
5.2.1 Quais fatores e como eles influenciam a recomendação	p. 75
5.3 Discussões Sobre o Capítulo 5	p. 82
6 Considerações Finais	p. 84
Referências Bibliográficas	p. 87

Capítulo 1

Introdução

A diversidade de serviços presentes na *web* tornou mais acessível a produção e o compartilhamento de conteúdo multimídia autoral. Uma consequência direta desse novo cenário foi o aumento da quantidade de dados produzido, armazenado e distribuído pelas aplicações. O Youtube (Youtube, 2016), a plataforma de produção e distribuição de vídeos de curta duração mais conhecida da Internet, registra uma carga de *upload* de mais de 300 horas de vídeo por minuto.

Essa nova realidade demanda também que técnicas criadas para organizar e recuperar conteúdo multimídia sejam aprimoradas. Uma técnica, chamada de etiquetagem (do inglês *tagging*), tem sido objeto de estudo recentemente. Segundo (WANG et al., 2012) etiquetagem é o processo de atribuição de um conjunto de etiquetas (do inglês *tags*) a um objeto com o intuito de descrever de forma resumida o seu conteúdo. Esse processo tem sido aplicado em texto, vídeo e imagem como forma de expressar seu conteúdo nos níveis sintático e semântico. Em (FIGUEIREDO et al., 2009) afirma-se que dentre os atributos de um objeto, a etiqueta é o atributo que possui a melhor capacidade descritiva, devido ao poder de resumo do conteúdo associado ao objeto.

Segundo (WANG et al., 2012) o processo de etiquetagem pode ser realizado de forma manual, automática ou assistiva. Na etiquetagem manual a atribuição de etiquetas ocorre na ausência de procedimento computacional, demandando do responsável um grande esforço criativo e vasto vocabulário. Em geral, o processo é repetitivo e exaustivo, resultando em descrições de baixa qualidade. A etiquetagem automática atribui o conjunto de etiquetas ao um objeto sem interferência humana direta. Usa-se técnicas de aprendizado de máquina para aperfeiçoar o processo, sendo a ausência de contexto o principal desafio. A etiquetagem assistiva explora a interação entre procedimentos manuais e computacionais,

tendo como objetivo explorar as melhores características presentes na etiquetagem manual e na etiquetagem automática, ou seja a noção de contexto e o vasto vocabulário.

A presente proposta baseia-se no processo de etiquetagem assistiva, especificamente em recomendação de etiquetas. A recomendação de etiquetas disponibiliza palavras chaves que mais caracterizam um objeto, usufruindo da intervenção humana durante o processo. Os estudos nessa área têm usado métodos associativos que exploram padrões de coocorrência para gerar etiquetas candidatas a recomendação. Nessa abordagem o desafio é a adaptação do processo de recomendação ao fato do objeto alvo possuir poucas informações, reconhecido como estado de *cold start*, o que pode resultar em recomendações com baixa relevância requerendo do interventor humano um esforço muito próximo daquele empreendido no processo manual.

Este trabalho apresenta um estudo exploratório sobre a integração de um método associativo de recomendação de etiquetas com uma técnica que busca os conceitos mais significativos no conjunto dos termos candidatos a recomendação considerando-se os impactos na qualidade do conjunto de etiquetas geradas. O estudo considera a necessidade da minimização dos ruídos nos dados gerados pela intervenção humana em um estado de *cold start* e tem como estudo de caso uma aplicação de distribuição de vídeos de curta duração, onde cada vídeo possui meta dados associado, e.g título e descrição.

1.1 Problema

No contexto da aplicação considerada neste trabalho, um dos maiores desafios para as técnicas de recuperação de informação, além do tamanho das coleções de dados e das altas taxas de renovação de conteúdo, é a qualidade dos metadados associados aos objetos gerados pelos usuários. A recomendação de etiquetas tem sido utilizada em diversas aplicações visando melhorar a qualidade desses metadados. Entretanto, a introdução do fator humano torna o processo dependente dos conhecimentos de quem o opera, impõem-se assim a necessidade de redução do esforço empregado ao mesmo tempo que é preciso preservar a qualidade das etiquetas recomendadas.

Outra questão a ser considerada é como recomendar etiquetas para objetos em estado de *cold start*, i.e. com reduzidas fontes de informação associada ao objeto. Objetos com essa característica estão presentes nas estatísticas coletadas em diversas aplicações da Web 2.0. Em (FIGUEIREDO et al., 2009) mostra-se que tais objetos representam 18% de todos os objetos encontrados em aplicações populares.

1.2 Motivação

Segundo estatísticas reportadas pela Sandvine (Sandvine, 2016), no segundo semestre de 2016, aplicações da categoria entretenimento em tempo real são responsáveis por 46% de todo o tráfego de rede produzido na América Latina, ver Figura 1.1.

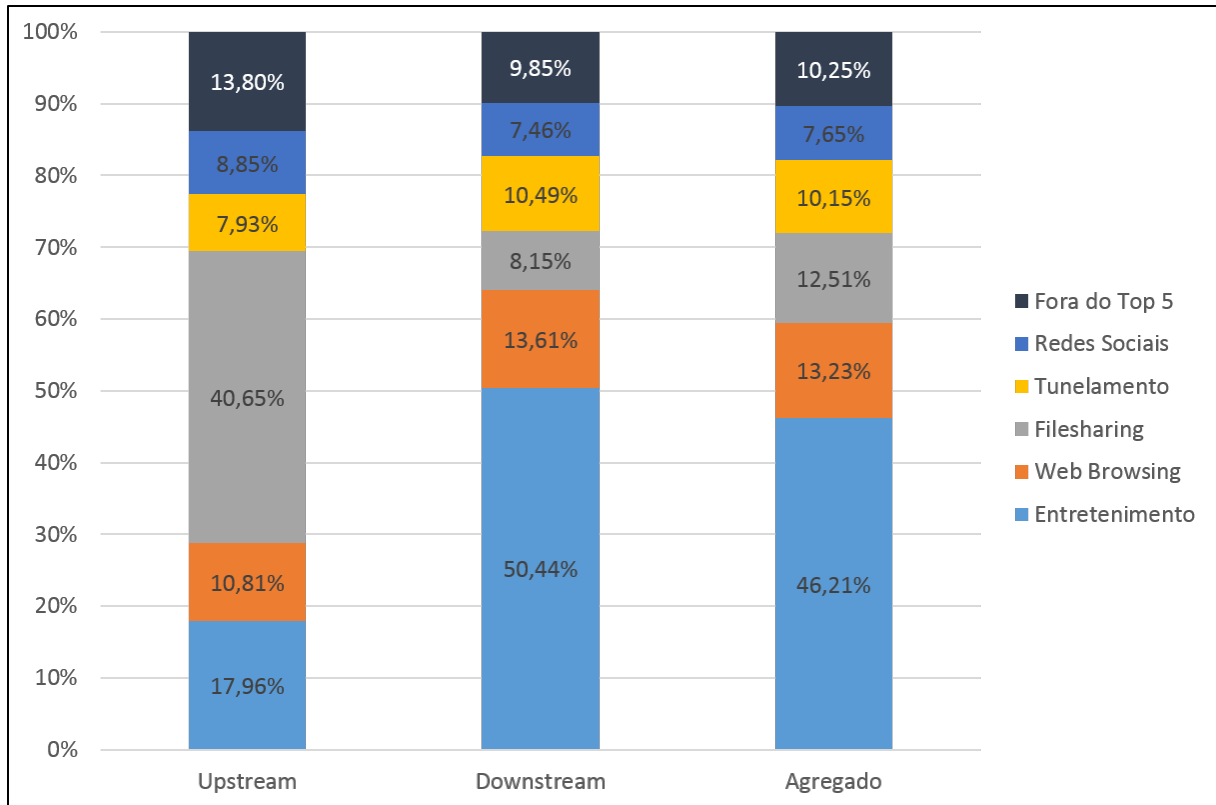


Figura 1.1: Tráfego da rede na América Latina em horário de pico

Entre as dez aplicações que mais geraram tráfego na Rede, metade são aplicações relacionadas a vídeos, com a participação destacada da Netflix, que passou de 6.6% no primeiro semestre de 2015 para 8.31% em 2016, cerca de 25% de aumento de tráfego, ver Figura 1.2. Já o Youtube, continua sendo a aplicação que mais gera tráfego na Rede, além de ser umas das aplicações que mais recebem conteúdo na *web*.

O cenário das aplicações de distribuição de vídeos, descrito anteriormente, associado ao alto custo computacional gerado pela extração de informações diretamente dos vídeos, sugerem a etiquetagem como recurso valioso a ser utilizado para a classificação, busca e recomendação desse conteúdo. Ao se tornar peça chave nesse processo, o avanço no estado da arte das técnicas de recomendação de etiquetas, devem promover uma melhor descrição do conteúdo dos vídeos, impactando positivamente na eficiência dos serviços de recomendação de conteúdo que se instalam naquelas aplicações.

	Upstream		Downstream		Agregado	
Rank	Aplicação	tráfego	Aplicação	tráfego	Aplicação	tráfego
1	BitTorrent	30.03%	Youtube	28.48%	Youtube	25.91%
2	Youtube	9.30%	HTTP - outros	11.66%	HTTP – outros	11.12%
3	HTTP - outros	7.59%	SSL - outros	9.76%	BitTorrent	10.06%
4	Facebook	6.72%	Netflix	8.31%	SSL – outros	9.28%
5	SSL - outros	6.19%	BitTorrent	6.96%	Netflix	7.45%
6	Ares	5.27%	Facebook	5.10%	Facebook	5.32%
7	Skype	2.53%	MPEG - outros	2.28%	MPEG - outros	2.10%
8	Netflix	1.97%	RTMP	1.79%	RTMP	1.66%
9	Dropbox	1.16%	Google Market	1.69%	Google Market	1.52%
10	MPEG - outros	0.92%	Flash Video	1.60%	Flash Video	1.46%
		71.69%		77.63%		75.87%

Figura 1.2: As 10 aplicações que mais geram tráfego no horário de pico na América Latina

1.3 Objetivo

Esta dissertação é motivada pelo seguinte objetivo:

Propor e avaliar uma nova estratégia de recomendação associativa de etiquetas que explora grafos de contexto, para melhorar a qualidade da recomendação de etiquetas em vídeos de conteúdo esportivo em estado de *cold start*. A avaliação desta estratégia deve esta de acordo com as métricas definidas.

Para garantir o objetivo, definiu-se as seguintes atividades:

- Definir e implementar os módulos do método associativo pré-definido e da estratégia de recomendação associativa de etiquetas que explora grafos de contexto;
- Definir métrica de relevância de etiquetas a ser considerada para viabilizar a comparação de desempenho entre as estratégias com os cenários diferentes definidos em uma avaliação quantitativa.
- Definir os fatores a serem considerados influentes nas recomendações em uma avaliação qualitativa das estratégias.
- Elaborar uma dinâmica de avaliação das recomendações de etiquetas para uma aplicação de vídeos;

1.4 Principais Contribuições

A partir do desenvolvimento dos objetivos definidos neste trabalho foi possível realizar as seguintes contribuições:

1. Extração das principais características dos métodos de recomendação associativa e de outros métodos que abordam o problema de *cold start* presentes na literatura;
2. Elaboração de um conjunto de scripts e módulos para a coleta de informações sobre vídeos e recomendações das etiquetas;
3. Apresentação de um estudo quantitativo do desempenho das estratégias de recomendação associativa de etiquetas estudadas; e;
4. Estudo qualitativo das recomendações produzidas pelas estratégias, e a constatação de que a estratégia proposta apresenta melhoria de desempenho em vídeos que possuem até uma quantidade mediana de etiquetas.

1.5 Estrutura da Dissertação

Esta dissertação está organizada em cinco capítulos. No capítulo 2 apresenta-se o referencial teórico, necessário para o entendimento mínimo do trabalho, e uma descrição do estado da arte das estratégias de recomendação associativa de etiquetas nos estados sem e com *cold start* e de outras estratégias de recomendação de etiquetas que abordam o problema de *cold start*, através dos quais foram identificadas características que levaram a produção de uma tabela comparativa dessas estratégias.

No capítulo 3 apresentam-se o método base escolhido para este trabalho, as bases de dados, a metodologia de avaliação, e o ambiente de experimentação utilizados, bem como a caracterização dos cenários. Por fim, apresenta-se uma análise do desempenho do método base, Sum^+wTS .

No capítulo 4 apresentam-se a circunstância na qual o grafo de contexto é utilizado e a estratégia de recomendação associativa usando os grafos de contexto proposta. Além disso, apresenta-se um estudo comparativo entre a estratégia desenvolvida e o método base escolhido.

No capítulo 5 apresentam-se o ambiente de experimentação desenvolvido para a avaliação qualitativa e o estudo das recomendações geradas pela estratégia proposta nesse tipo de avaliação.

Por fim, no capítulo 6 apresentam-se as considerações finais desta dissertação, suas limitações, conclusões à respeito dos resultados obtidos, bem como os trabalhos futuros.

Capítulo 2

Conceitos e Trabalhos Relacionados

Neste capítulo apresentam-se os principais conceitos relacionados a recomendação de etiquetas e os métodos associativos, necessários ao entendimento do contexto em que este trabalho está inserido. Especificamente, aborda-se a recomendação de etiquetas e seus paradigmas, os métodos associativos e o problema de *cold start*. Além disso, apresentam-se as estratégias de recomendação associativa, bem como outros métodos que exploram o problema de *cold start* presentes na literatura.

O levantamento bibliográfico foi conduzido a partir de um conjunto de questões que relacionam os métodos de recomendação de etiquetas, principalmente os métodos associativos ao problema de *cold start* e as formas de avaliação do desempenho das recomendações. As questões postas foram estas: Quais são as principais estratégias de recomendação associativa de etiquetas que avaliam o processo de etiquetagem em cenário de *cold start*? A utilização de grafo de contexto em uma estratégia para técnicas de recomendação associativa de etiquetas é robusta no estado de *cold start*? Quais são os métodos existentes para avaliação da qualidade da recomendação de etiquetas nestas técnicas? O conjunto de artigos catalogado a partir das consultas as bases de dados são discutidos a seguir.

Este capítulo organiza-se da seguinte forma: Na Seção 2.1 apresentam-se o conceito e a classificação de recomendação de etiquetas presentes na literatura, além dos conceitos de método associativo e o problema de *cold start*. Na Seção 2.2 apresentam-se os trabalhos relacionados aos métodos associativos de recomendação de etiquetas que exploram etiquetas previamente associadas a um objeto. Na Seção 2.3 apresentam-se os trabalhos relacionados aos métodos associativos que abrangem o problema de *cold start*. Na Seção 2.4 apresentam-se outros métodos de recomendação de etiquetas que abordam o problema de *cold start*. E por fim, na Seção 2.5 estão as considerações finais deste

capítulo.

2.1 Conceitos

2.1.1 Recomendação de etiquetas

Segundo (WANG et al., 2012), recomendação de etiquetas é uma forma de etiquetagem assistiva, em que sugere-se um conjunto de etiquetas candidatas ao responsável pelo processo de etiquetagem. O responsável pelo processo deverá selecionar, do conjunto de etiquetas candidatas, os termos que mais descrevem o objeto a ser marcado, visando a melhoria da eficiência da acurácia das sugestões. A tarefa de atribuição de etiquetas requer um alto esforço em sua realização, e geralmente os usuários provêm etiquetas com baixa qualidade, termos com erros de ortografia ou descontextualizadas.

Em (LIPCZAK; MILIOS, 2011) os sistemas de recomendação de etiquetas são agrupadas em três categorias: baseado em grafos, baseado em conteúdo e sistemas híbridos. A recomendação de etiquetas baseada em grafos utiliza as relações de coocorrência entre usuários, recursos e etiquetas. Essas relações representam a *folksonomy*, i.e. classificação de conteúdo realizada pelos usuários do sistema, e o grafo formado por elas é usado para sugerir termos relevantes, utilizando algoritmos de ranqueamento. A recomendação baseada em conteúdo utiliza somente os metadados textuais relacionados ao conteúdo do objeto alvo. E por fim, os sistemas híbridos combinam os dois tipos de entrada das categorias anteriores.

Em (BELÉM et al., 2014) aborda-se o problema de recomendação de etiquetas sob duas perspectivas. A primeira, chamada de centrada no objeto, visa sugerir etiquetas relevantes para um determinado objeto. Enquanto que a segunda perspectiva, centrada no objeto e no usuário, tem como finalidade a realização de recomendação de etiquetas personalizadas para um par objeto-usuário. Em comparação com a classificação realizada em (LIPCZAK; MILIOS, 2011), a primeira e a terceira categorias descritas estão localizadas na segunda perspectiva no trabalho descrito em (BELÉM et al., 2014) e a segunda categoria está localizada na primeira perspectiva, sendo os métodos contidos nesta perspectiva como base para a pesquisa desenvolvida neste trabalho.

Em (BELÉM; ALMEIDA; GONÇALVES, 2017) apresenta-se uma taxonomia dos métodos de recomendação de etiquetas, classificando-os de acordo com os seguintes fatores: o tipo de alvo das recomendações, seus objetivos, fonte de dados explorados, e técnicas utilizadas. Além disso, prover-se uma análise crítica desses métodos, pontuando as suas

vantagens e desvantagens. Descreve-se os principais desafios em aberto na área como problemas de semântica, *spamming*, esparsidade, *cold start* e avaliações dos métodos.

2.1.2 Método associativo

Os métodos associativos de recomendação de etiquetas exploram padrões de coocorrência obtidos das etiquetas previamente associadas ao objeto o e de uma coleção de objetos de treino (BELÉM et al., 2011). Os padrões de coocorrência são aprendidos por meio da exploração das regras de associação adquiridas de um conjunto de treino $\mathcal{D} = \{\langle \mathcal{L}_d, \mathcal{F}_d \rangle\}$, onde $\mathcal{L}_d$ ($\mathcal{L}_d \neq \emptyset$) é o conjunto composto de todas as etiquetas e $\mathcal{F}_d$ é o conjunto composto dos outros atributos textuais do objeto $d \in \mathcal{D}$.

A regra de associação é uma implicação do tipo $X \rightarrow y$, onde o antecedente X é um conjunto de etiquetas e o consequente y é um termo candidato à recomendação. Os padrões obtidos dessas regras são utilizados para recomendar etiquetas para objetos em um conjunto de teste $\mathcal{O} = \{\langle \mathcal{L}_o, \mathcal{F}_o \rangle\}$. Em outras palavras, estes métodos expandem o conjunto $\mathcal{L}_o$ inicial do objeto o usando os termos de outros atributos textuais do objeto e os termos que estão presentes no conjunto de treino e que frequentemente coocorrem com as etiquetas.

2.1.3 Cold start

A maioria dos métodos associativos considera que o objeto alvo de recomendação possui etiquetas previamente associadas. Entretanto, existem situações em que essa condição não é verificada, caracterizando o problema de *cold start*. Esse problema foi definido em (SCHEIN et al., 2002) e é descrito pelo estado em que não há informações sobre as preferências do usuário.

No contexto de métodos associativos de recomendação de etiquetas, o problema de *cold start* é caracterizado por um estado em que o objeto alvo da recomendação não possui nenhuma etiqueta previamente associada, mas pode conter termos em seus outros atributos. Tal cenário corresponde à adição de um novo objeto ao sistema, sendo necessário recomendar as primeiras etiquetas que o descreva.

2.2 Métodos associativos

No estudo realizado em (SIGURBJÖRNSSON; Van Zwol, 2008) propõe-se um método de recomendação de etiquetas que explora o conhecimento coletivo da comunidade *Flickr*. Esse método usa as relações entre as etiquetas, por meio de métricas globais de coocorrência. Quatro estratégias são apresentadas, sendo compostas pela extração de estatísticas de coocorrência de etiquetas, juntamente com duas estratégias de agregação e uma função de promoção de etiquetas.

Além disso, os autores apresentam uma análise do comportamento dos usuários ao realizarem o processo de etiquetagem considerando as seguintes questões: como os usuários marcam as fotos? e que tipo de etiquetas eles fornecem?. Por meio da geração de um modelo de distribuição de frequência e do mapeamento das etiquetas da base, observa-se que maior parte das etiquetas geradas pelos usuários se refere a quem, o que é, e quando a foto foi tirada. Os resultados mostram que o método proposto mantém o desempenho com a evolução do vocabulário.

Em (MENEZES et al., 2010) propõe-se um método de recomendação de etiquetas baseado em coocorrência chamado LATRE. Esse método extrai regras de associação a partir dos dados em conjunto de treino, de acordo com um conjunto inicial de etiquetas atribuídas a um objeto. O método prevê o espaço de busca por regras, de acordo com a qualidade das informações dos objetos de teste, permitindo a extração de regras de associação mais elaboradas.

No LATRE, cada regra extraída é interpretada como um voto para a etiqueta candidata. Após a extração, os votos são somados e as etiquetas são ordenadas segundo sua pontuação. Por fim, o método calibra as pontuações para corrigir possíveis distorções da lista de etiquetas ranqueada. O método LATRE recomenda etiquetas com precisão maior em todos os subconjuntos de dados, em comparação com o seu *baseline*. Os ganhos alcançados estão no intervalo de 6% a 46,7% dependendo do conjunto de dados utilizado.

Em (BELÉM et al., 2011) estende-se o trabalho desenvolvido em (SIGURBJÖRNSSON; Van Zwol, 2008) com a proposta de dez novas heurísticas. Essas heurísticas exploram as três dimensões do problema de recomendação de etiquetas: a coocorrência de termos com etiquetas previamente associadas a um determinado objeto, termos extraídos de múltiplos recursos textuais de um objeto, e métodos que calculam a relevância das etiquetas.

Os autores exploram também o uso de duas técnicas de *Learning To Rank* (L2R):

RankSVM e Programação Genética. Essas técnicas são usadas na geração de funções de ranqueamento que usam múltiplas métricas, melhorando a estimativa da relevância das etiquetas de um dado objeto. Assim, investiga-se os benefícios da aplicação dessas técnicas na recomendação de etiquetas.

Os resultados em (BELÉM et al., 2011) mostram que as heurísticas propostas proporcionam ganhos em relação as técnicas estado-da-arte. A melhor heurística promove ganhos de até 40% em precisão, sobre o melhor *baseline* em qualquer cenário avaliado. As técnicas L2R apresentam resultados competitivos em relação aos métodos associativos, além de serem flexíveis, permitindo a inclusão de novas métricas e a extensão do escopo do problema.

Em (OLIVEIRA et al., 2012) é proposto um método para expansão de termos extraídos de atributos textuais de um objeto alvo de recomendação de etiquetas. Esse método filtra, pontua e ranqueia termos candidatos extraídos de artigos da Wikipédia. O método de expansão proposto é aplicado nos métodos de recomendação de etiquetas apresentados em (BELÉM et al., 2011), que explora meta-heurísticas para estimar a capacidade descritiva de um dado termo.

A racionalidade do método de expansão é que os termos mais relevantes de um artigo da Wikipédia, segundo o método de recomendação de etiquetas, também são os mais relevantes para a expansão da consulta, i.e. termos extraídos de atributos textuais do objeto, desde que o artigo e a consulta estejam relacionadas semanticamente.

Os artigos do Wikipédia são as fontes primárias utilizadas neste processo de expansão. Dado uma consulta e o artigo relacionado, o objetivo do método é a produção de um conjunto de termos ranqueados. Esses termos podem ser utilizados para a expansão da consulta e consequentemente em recomendação de etiquetas.

É importante notar que esses métodos foram avaliados apenas em cenários em que os objetos alvo de recomendação possuem etiquetas previamente associadas, ou seja, sem o impacto do problema de *cold start*.

2.3 Métodos associativos e o *cold start*

No trabalho descrito em (LIPCZAK et al., 2009) é apresentado um sistema de recomendação de etiquetas que compila um conjunto de etiquetas oriundas do título do objeto, do histórico de etiquetagem de objetos similares, e do perfil do usuário. O sistema inicia a recomendação gerando um conjunto de etiquetas T oriundas de uma fonte textual

associada ao objeto, na proposta usa-se o título. O conjunto T é estendido com etiquetas que coocorrem na base do sistema. Em seguida, etiquetas relacionadas ao perfil do usuário no sistema são extraídas para formar um conjunto P . A interseção dos conjuntos T e P gera o conjunto de etiquetas F que relaciona o objeto ao usuário, na tentativa de estender o conjunto encontrando mais etiquetas relacionadas com o perfil do usuário. Por fim, o sistema extrai etiquetas do conteúdo do objeto e as une com as etiquetas em F , gerando a recomendação final.

Os autores avaliam a qualidade de três fontes básicas de etiquetas: termos do título do objeto, etiquetas associadas a objetos similares por outros usuários e etiquetas do perfil do usuário. Os resultados consideram que o título do objeto como um ponto de partida natural para o processo de recomendação de etiquetas.

Em (GRAHAM; CAVERLEE, 2008) apresenta-se um *framework* de etiquetagem iterativo chamado *Plurality*. Este *framework* incorpora o *feedback* do usuário para fornecer sugestões de etiquetas para objetos que não possuem etiquetas associadas previamente. O *feedback* do usuário é utilizado no desenvolvimento e avaliação de três modelos: *tag-based*, *Term-Based* e *Tag Co-location*. Esses modelos são baseados no conceito de *relevance feedback* e exploram diferentes hipóteses sobre o melhor caminho para capturar a opinião do usuário, e as possibilidades do espaço de etiquetas existentes para a recomendação.

O processo iterativo para a identificação de etiquetas, mantém os usuários interagindo com o sistema, guiando o processo de seleção das etiquetas.

Os resultados mostram que os modelos *Tag Feedback* e *Term Feedback* são um processo de refinamento, que altera os objetos relacionados e conseqüentemente a lista de etiquetas resultante. O modelo *Tag Co-location* requer menor esforço por parte dos usuários no processo de etiquetagem e obtém a maior razão de boas etiquetas recomendadas para os usuários.

No estudo realizado em (MARTINS et al., 2016) é proposto uma avaliação do impacto do problema de *cold start* sobre os principais métodos associativos de recomendação de etiquetas. Além disso, apresenta-se uma nova estratégia de recomendação de etiquetas que explora tanto o *feedback* positivo quanto o *feedback* negativo dos usuários aliado a aplicação de técnicas de programação genética e *learning-to-rank*.

A estratégia de recomendação incorpora as preferências por etiquetas específicas manifestada pelos usuários durante o processo. Assim, as etiquetas definidas como relevantes pelo usuário são usadas como entrada para métodos associativos de etiquetas para posterior recomendação. As etiquetas que não são consideradas relevantes são inseridas em

uma *blacklist*, para que não sejam recomendadas nas próximas iterações.

Os resultados revelam que a acurácia dos métodos associativos de recomendação de etiquetas, considerados estado da arte, é bastante prejudicada, em média a perda de precisão é de 40% nas coleções do YahooVideo e do bibsonomy e de 28% no Youtube. A estratégia baseada em *relevance feedback* obteve ganhos de até 45% em comparação aos métodos associativos em *cold start*, e 39% em cenário onde o objeto possui etiquetas associadas previamente.

2.4 Outros métodos de recomendação de etiquetas e o *cold start*

Em (PREISACH; MARINHO; SCHMIDT-THIEME, 2010) apresenta-se um algoritmo semi-supervisionado baseado em grafos para realizar recomendações personalizadas. O grafo é produzido através da observação dos perfis de usuário das aplicações: bibsonomy, e LastFm, que geralmente contêm *posts* não marcados. Esses *posts* têm suas informações estruturais exploradas para melhorar as recomendações, principalmente em cenários em que poucos *posts* marcados estão disponíveis.

Em (NESS et al., 2009) propõe-se um método baseado na análise do áudio do recurso (música) em conjunto com o sistema de recomendação de etiquetas para conteúdo musical em estado de *cold start*. Entretanto, o método utiliza algoritmos especializados em cada tipo de objeto para a realização da análise de conteúdo, tornando-se um método de alta complexidade.

No trabalho descrito em (RAFAILIDIS et al., 2014) apresenta-se um método de recomendação de etiquetas personalizado baseado em etiquetagem social. Este método realiza a propagação de etiquetas baseado em conteúdo para tratar o problema de *cold start* que ocorrem em sistemas de etiquetagem social. O método explora o conteúdo dos itens, o histórico de atribuições de etiquetas dos usuários por meio de um mecanismo de *feedback*.

Em (HONG; KIM; LEE, 2017) apresenta-se uma técnica de recomendação de etiquetas semântica para imagens que explora os termos associados que são semanticamente similares ou relacionados entre si usando as ligações entre os artigos da Wikipedia. A técnica consiste em duas etapas. Na primeira criam-se grafos de relacionamento dos termos significativos de cada artigo da Wikipedia, de modo *offline*. Estes termos são reorganizados de acordo com o valor atribuído por algoritmos de ranqueamento baseado em análise de *links* e então formam o conjunto de termos associados a dado artigo. Na segunda etapa

um usuário realiza o *upload* de uma imagem, coleta-se imagens semelhantes e suas etiquetas que tornam-se etiquetas candidatas. Em seguida recomenda-se os termos associados relacionados as etiquetas candidatas. A técnica proposta apresenta ganhos de 14% de acurácia em relação a outros trabalhos de recomendação de etiquetas para imagens que utilizam bases de conhecimento.

Os métodos de recomendação de etiquetas apresentados nesta seção não são caracterizados como métodos associativos, que são o foco deste trabalho, entretanto, tratam o problema de *cold start* no processo de recomendação de etiquetas.

Na Tabela 2.4 mostra-se uma comparação entre os métodos de recomendação de etiquetas listados neste capítulo. A comparação é realizada observando os seguintes aspectos: a fonte na qual os termos são extraídos, a estratégia de recomendação utilizada e se o método abrange o estado de *cold start*.

Tabela 2.1: Comparação das características dos métodos de recomendação de etiquetas.

Trabalho	Fonte de termos	Estratégia	Cold Start
(SIGURBJÖRNSSON; Zwol, 2008)	Etiquetas previamente associadas.	Extração de estatísticas de coocorrência de etiquetas, juntamente com duas estratégias de agregação e uma função de promoção de etiquetas.	Não
(MENEZES et al., 2010)	Etiquetas previamente associadas.	regras de associação (coocorrência) + <i>demand-driven</i>	Não
(BELEM et al., 2011)	Etiquetas previamente associadas e termos oriundos dos títulos e descrição dos vídeos	10 estratégias de recomendação de etiquetas que incluem novas métricas de relevância e técnicas de <i>Learning-to-Ranking</i> (L2R).	Não
(OLIVEIRA et al., 2012)	Atributos textuais do objeto	expansão de termos + métricas heurísticas de relevância	Não
(LIPCZAK et al., 2009)	Título do objeto + histórico de etiquetagem de objetos similares e do usuário	Merge dos conjuntos de etiquetas oriundas de seis recomendadores de etiquetas	Sim
(GRAHAM; CAVERLEE, 2008)	Etiquetas de outros objetos similares da coleção	Modelos de <i>feedback</i> (positivo)	Sim
(MARTINS et al., 2013) e (MARTINS et al., 2016)	Título e descrição do objeto	<i>Relevance Feedback</i> (positivo e negativo) e técnicas de programação genética e learning-to-rank	Sim
(PREISACH; MARINHO; SCHMIDT-THIEME, 2010)	Informação estrutural de posts não marcados	Método relacional semi-supervisionado	Sim
(NESS et al., 2009)	Recursos do áudio e palavras associadas ao áudio	Classificador SVM + modelo probabilístico	Sim
(RAFAILIDIS et al., 2014)	Etiquetas associadas pelos usuários	Mecanismo de relevance feedback + propagação de etiquetas para itens semelhantes	Sim
(HONG; KIM; LEE, 2017)	Etiquetas de imagens similares na base e da Wikipedia	propagação de etiquetas de imagens similares com associação à termos oriundos de ligações de artigos (semântica similar) na wikipedia	Sim

2.5 Considerações Finais do Capítulo 2

Neste capítulo foram apresentados os conceitos fundamentais que o envolvem a recomendação de etiquetas, os métodos associativos de recomendação de etiquetas e o problema de *cold start*.

Abordou-se a localização do paradigma de recomendação de etiquetas entre as abordagens de etiquetagem existentes, bem como a classificação dos sistemas de recomendação de etiquetas e suas perspectivas. Em seguida, foram apresentados os principais trabalhos que envolvem os métodos associativos que abarcam meios que extraem desde o conhecimento da comunidade, heurísticas que englobam métricas de relevância de etiquetas e uso de atributos textuais até o emprego de técnicas de programação genética e *learning-to-rank*.

Existem diversos métodos que utilizam a coocorrência, porém a maioria não considera o problema de *cold start*. Os trabalhos que abordam este aspecto mostram que a falta de etiquetas previamente associadas aos objetos impactam significativamente o desempenho destes métodos. E a forma encontrada para contornar este impacto volta-se para uso dos atributos textuais do objeto, informações do perfil e o conceito de *feedback* dos usuários, sendo este último mais explorado atualmente.

Os métodos que utilizam o conceito de *feedback* apresentam um dos melhores desempenhos dentre as estratégias de recomendação associativa de etiquetas, porém estes métodos são dependentes da participação dos usuários, principalmente em aplicações colaborativas. Dentre as demais técnicas que abordam o *cold start* destaca-se o uso de análise da composição dos objetos alvos da recomendação, exploração do perfil e do histórico de etiquetagem dos usuários, uso de grafos e o conhecimento de bases de conhecimento como a Wikipédia. Por fim, apresentou-se um quadro comparativo reunindo as principais características atribuídas as estratégias encontradas nos trabalhos listados.

Capítulo 3

Método associativo de recomendação de etiquetas Sum^+wTS

Neste capítulo, apresenta-se o método Sum^+wTS , considerado um dos métodos associativos de recomendação de etiquetas estado da arte, sendo escolhido como base de comparação neste trabalho. Além disso, apresentam-se o arcabouço necessário para a realização da avaliação automática deste método e os resultados de uma análise experimental exploratória que avaliou o seu desempenho em uma coleção de vídeos esportivos. Nesta parte da pesquisa avalia-se o Sum^+wTS em estado de *cold start* assim como no estudo conduzido em (MARTINS et al., 2013), entretanto utiliza-se uma coleção de dados mais específica.

Especificamente, na primeira parte do estudo valida-se a implementação do método Sum^+wTS tendo-se os resultados em (BELÉM et al., 2011) e (MARTINS et al., 2013) como bases de comparação. Na segunda parte verifica-se o comportamento do método quando a base Youtube_Sports é utilizada. A metodologia de avaliação e os cenários definidos na Seção 3.3 foram utilizados nesse estudo.

Este capítulo organiza-se da seguinte forma: Na Seção 3.1 apresenta-se a composição do método Sum^+wTS . Na Seção 3.2 apresenta-se as coleções de dados utilizada nos experimentos. Na Seção 3.3 apresenta-se a metodologia de avaliação automática utilizada, com os cenários de recomendação definidos para os experimentos e a métrica utilizada para avaliação do desempenho do método experimentado. Na Seção 3.4 apresenta-se o ambiente utilizado para os experimentos. Na Seção 3.5 apresenta-se a análise dos resultados obtidos nos experimentos de validação da implementação do método. E por fim, na Seção 3.6 estão as considerações finais deste capítulo.

3.1 Sum^+wTS

O método Sum^+wTS , proposto em (BELÉM et al., 2011), explora conjuntamente as etiquetas previamente associadas ao objeto, os atributos textuais e métricas de relevância de etiquetas. O Sum^+wTS sendo uma conjunção do método Sum^+ e da heurística wTS .

O Sum^+ realiza a recomendação baseado em duas métricas: *Sum* e *Estabilidade*. A métrica *Sum* estima a relevância de um termo candidato c como etiqueta para o objeto o pela soma das confianças (θ) de todas as regras de associação contendo c como consequente, sendo que o antecedente da regra de associação deve está contido no conjunto de etiquetas previamente associadas ao objeto. A Equação 3.1 apresenta o método *Sum*.

$$Sum(c, o, l) = \sum_{X \subseteq \mathcal{L}_o} \theta(X \rightarrow c), \quad (X \rightarrow c) \in R, |X| \leq l \quad (3.1)$$

onde R é um conjunto de regras de associação computadas previamente sobre um conjunto de treino $\mathcal{D}$, e l é a quantidade de etiquetas permitidas no antecedente de uma regra de associação.

O *Sum* retorna novas etiquetas de uma base de treino usando a coocorrência entre os termos. Entretanto, existem termos que apesar de coocorrerem com as etiquetas do objeto, podem apresentar uma frequência nos extremos, i.e são termos com baixa ocorrência ou com ocorrência alta. Essa não é uma característica de um bom termo, pois são gerais e tem um baixo poder descritivo do conteúdo do objeto. Assim, a métrica *Estabilidade* foi integrada ao método, dando mais ênfase para os termos candidatos que ocorrem com uma frequência intermediária. A métrica *Estabilidade* (Est) é dada como segue:

$$Estabilidade(c, k_s, l) = \frac{k_s}{k_s + |k_s - \log(f_c^{tag})|} \quad (3.2)$$

onde k_s representa a frequência ideal de um termo, e f_c^{tag} é o número de objetos no conjunto de treino $\mathcal{D}$ que contêm o termo c como etiqueta.

Na definição do método Sum^+ , o conceito da *Estabilidade* é empregado, e a Equação 3.1 é redefinida como segue:

$$Sum^+(c, o, k_i, k_c, k_r) = \sum_{i \subseteq \mathcal{L}_o} \theta(\{i\} \rightarrow c) \times Est(c, k_i) \times Est(c, k_c) \times Rank(c, i, k_r) \quad (3.3)$$

onde k_i , k_c e k_r são parametros de configuração; a função $Rank(c, i, k_r)$ é uma heurística que

pondera as etiquetas candidatas que estão no topo das listas de candidatas, suavizando a redução dos valores de coocorrência.

A função $Rank(c, i, k_r)$ é dada por $k_r / (k_r + p(c, o))$, onde $p(c, o)$ é a posição de c entre as etiquetas candidatas ordenadas pelo valor de Confiança.

Em (BELÉM et al., 2011) apresenta-se o método Sum^+wTS , que é uma melhoria do método Sum^+ , proporcionada pela integração da heurística wTS no processo de recomendação. A wTS mede o espalhamento do termo candidato c em um objeto o , ou seja, $TS(c, o)$ é dado pelo número de atributos textuais associado a o que contém c . A idéia é que quanto mais atributos contiverem um termo t_i , mais este termo estará relacionado ao conteúdo do objeto o . Assim uma etiqueta com um alto $TS(c, o)$ deve apresentar um bom poder descritivo para o conteúdo do objeto (FIGUEIREDO et al., 2009).

A heurística wTS captura o poder descritivo dos atributos textuais de um objeto, determinando o valor de um termo candidato por sua presença nesses atributos. No primeiro passo usa-se a $TS(t, o)$ média da instância de um atributo, e dessa forma estima-se o grau de relacionamento de uma instância de atributo com o conteúdo do objeto.

O passo seguinte é calcular o poder descritivo de um atributo em relação a coleção de objetos, que é dado pelo FIS médio. O FIS é a média dos valores TS sobre todos os termos de uma instância atributo de um objeto. Este valor médio é dado pela métrica Average Feature Spread, $AFS(F)$. A Equação 3.4 mostra o cálculo da wTS .

$$wTS(c, o) = \sum_{X \subseteq F_o^i \in F_o} j, \text{ onde } j = \begin{cases} AFS(F^i) & \text{se } c \in F_o^i \\ 0 & \text{caso contrário} \end{cases} \quad (3.4)$$

A conclusão desta integração que origina o Sum^+wTS produz um ranqueamento de termos candidatos baseados em coocorrência e métricas de qualidade de etiquetas. A Equação 3.5 apresenta o método Sum^+wTS .

$$Sum^+wTS(c, o, k_x, k_c, k_r, \alpha) = \alpha Sum^+(c, o, k_x, k_c, k_r) + (1 - \alpha) wTS(c, o) \quad (3.5)$$

onde k_x , k_c , k_r são parâmetros de ajuste. O parâmetro α pondera os valores obtidos pelas métricas de recomendação.

3.2 As Bases de dados

As estratégias de recomendação de etiquetas avaliadas neste trabalho utilizam amostras de três bases de dados: Youtube_collection_ufmg, Youtube_Sports_ufam, e Wikipédia. A base Youtube_collection_ufmg utilizada em (MARTINS et al., 2013) é formada por meta-dados de 5 milhões de vídeos, que foram coletados em agosto de 2009 através da técnica de amostragem bola de neve. Os meta-dados referem-se ao título, descrição, comentários e etiquetas dos vídeos, sendo considerado apenas atributos textuais em língua inglesa. Neste trabalho foi utilizada uma amostra disponível desta coleção, que será denominada de Youtube_ufmg. Esta amostra é composta por metadados de 150 mil vídeos escolhidos randomicamente e de diversas categorias desta base.

A base Youtube_Sports_ufam é formada por meta-dados que descrevem vídeos produzidos e publicados pelos usuários desta aplicação, que foram coletados no segundo semestre de 2016. Os metadados da base Youtube_Sports_ufam foram obtidos usando um coletor que implementa uma busca em largura no grafo $G(V,A)$ definido, onde V é o conjunto de vídeos coletados e A são as arestas construídas pela recomendação do Youtube. As coletas são iniciadas com um conjunto inicial de vídeos disponibilizados no *site* através dos seguintes *feeds*: Mais Qualificados, Mais Incluídos entre os favoritos, Mais Vistos, Mais Populares, Mais Discutidos, e Mais Respondidos. Esses *feeds* podem ser filtrados por país e categorias associadas aos vídeos. Para este trabalho, coletou-se apenas vídeos oferecidos aos brasileiros na categoria de esporte.

Cada vídeo obtido nos *feeds*, chamado de vídeo semente, é colocado em uma lista que é processada sequencialmente. Para cada vídeo nessa lista, um conjunto de vídeos relacionados é coletado do YouTube e adicionado na lista de vídeos para ser processado no futuro. A coleção de vídeos do YouTube é explorada em largura até que todos os vídeos alcançáveis a partir do conjunto de vídeos semente tenham sido visitados. Em resumo, o coletor implementa o conceito do método de coleta bola de neve.

Os dados da coleção foram gerados tanto pelo *site* quanto pelos usuários. Especificamente, para os dados gerados pelos usuários, o coletor obteve o título, a descrição do vídeo e as etiquetas atribuídas. Para os dados gerados pelo *site*, o coletor obteve o identificador do vídeo, a duração, contador de visualizações, a data de publicação, e os 25 primeiros vídeos relacionados, para cada vídeo inicialmente acessado. Atualmente, a coleção possui dados de 586551 vídeos, destes foram selecionados randomicamente 150 mil vídeos que formam a amostra Youtube_Sports, que é utilizada neste trabalho.

A Wikipédia é uma enciclopédia online colaborativa criada para a coleta e disseminação de conhecimento de forma livre. Os artigos podem ser criados e editados por qualquer colaborador desde que sigam as regras básicas estabelecidas pela comunidade, como a veracidade da informação e a relevância do tema. A Wikipédia disponibiliza cópia da coleção periodicamente, permitindo que o seu conhecimento seja explorado em diferentes propósitos. A coleção inclui o texto de todos os artigos, os metadados do artigo atual e suas revisões.

Os metadados dos artigos referem-se ao título, categoria, e os *links* existentes entre os artigos. Dos *links* pode-se extrair propriedades da estrutura das relações entre os artigos, como o sinonímia, polisemia e hierarquia. A sinonímia é obtida por meio das páginas de redirecionamento, que são páginas usadas para redirecionar o usuário a página que contém o artigo completo atual, i.e. quando existe diferentes modos de descrever o mesmo termo, como abreviaturas, ortografia múltipla, títulos em outros idiomas, palavras relacionadas e plurais. A polisemia é obtida por meio das páginas de desambiguação, que são páginas que apresentam os artigos que possuem a mesma grafia, mas tratam de conceitos distintos. Além das propriedades mencionadas os *links* também relacionam os artigos entre si de forma contextual.

A base de dados da Wikipédia utilizada foi a coleção de artigos da Wikipédia brasileira coletada da Wikimedia em Janeiro de 2016. Esta coleção contém 3,786,844 artigos, com cerca de 26 milhões de *links* entre eles. Destes artigos, cerca de 428 mil são páginas de redirecionamento. Além disso, a coleção possui informações sobre as categorias utilizadas e sua hierarquia, e uma lista com cerca de 2,1 milhões de títulos de artigos.

3.3 Metodologia de Avaliação

Para a avaliação das estratégias de recomendação de etiquetas foi adotada uma abordagem automática, similar a (BELÉM et al., 2011), onde um subconjunto das etiquetas associadas aos objetos do conjunto de teste é usado como gabarito para a recomendação, ou seja, são consideradas como etiquetas relevantes para o objeto, sendo posteriormente desconsideradas para o cálculo das métricas de relevância presentes em (BELÉM et al., 2011).

A avaliação foi realizada sobre os subconjuntos de etiquetas obtidos das amostras Youtube_ufmg, e Youtube_Sports descritas na Seção 3.2. Esses subconjuntos de etiquetas foram obtidos pela definição de dois cenários que caracterizam as amostras pela quantidade

de etiquetas presentes nos vídeos, buscando-se a simulação do desempenho dos métodos em vídeos com quantidade de etiquetas homogênea e heterogênea.

No primeiro cenário, divide-se a amostra em três subconjuntos de objetos baseados no número de etiquetas disponíveis em cada objeto, ver Figura 3.1. Os subconjuntos são formados e denominados como segue: *menor* (2 a 5 etiquetas), *médio* (6 a 9 etiquetas), e *maior* (acima de 9 etiquetas). Cada um desses subconjuntos possui uma amostra de 50 mil objetos selecionados aleatoriamente. Dessa forma, avaliou-se cada método separadamente para cada subconjunto, medindo-se a sua efetividade em amostras de dados com características semelhantes.

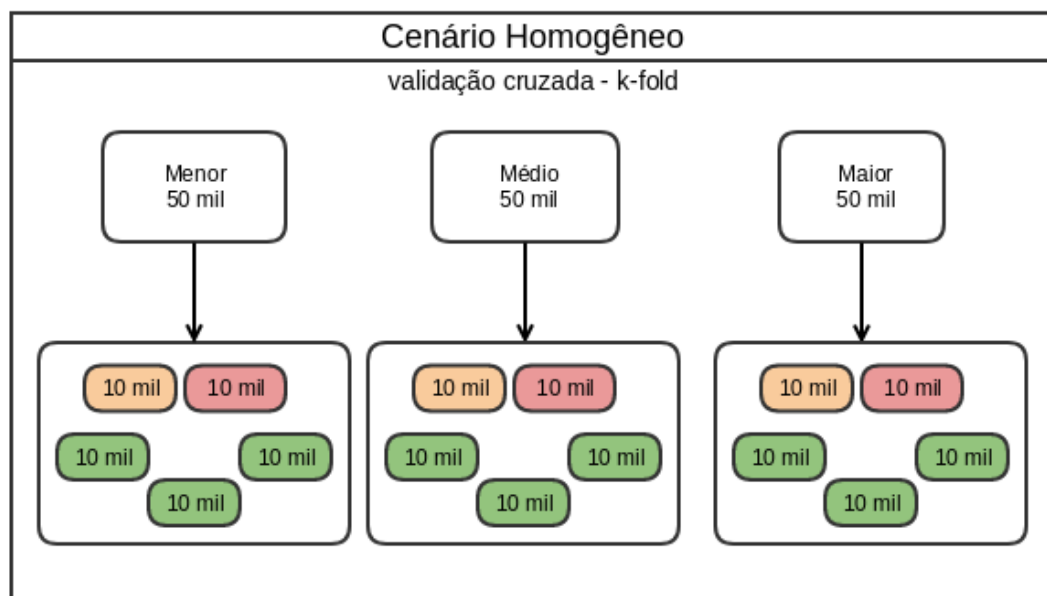


Figura 3.1: Cenário homogêneo.

No segundo cenário denominado de cenário heterogêneo, considera-se toda a amostra e analisa-se o comportamento dos métodos em uma subconjunto de etiquetas mais misto. Ou seja, a amostra deste cenário consiste na combinação de todos os três subconjuntos de etiquetas usados no primeiro cenário, ver Figura 3.2.

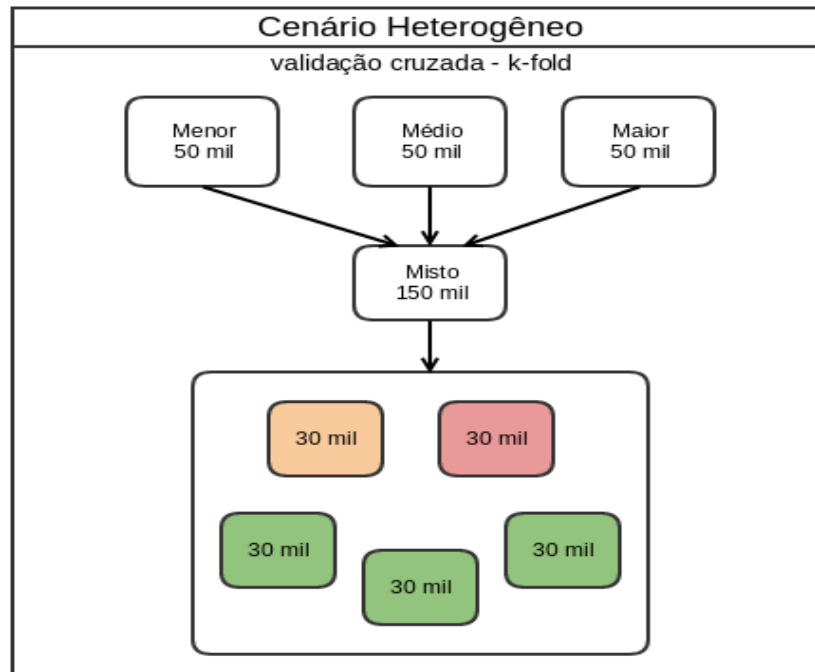


Figura 3.2: Cenário heterogêneo.

Assim para a avaliação do método cada subconjunto obtido de uma amostra é dividido em cinco partes iguais utilizadas em um procedimento experimental de validação cruzada (*cross-validation 5-fold*). Três partes são usadas como conjunto de treino, que por sua vez é utilizada para a extração das regras de associação e computação das métricas de relevância; uma parte é utilizada como conjunto de parametrização, e por fim, a última parte é utilizada como conjunto de teste O . Este processo é repetido cinco vezes, a cada rodada utiliza-se como conjunto de teste uma parte diferente. O resultado obtido é a média dos cinco desempenhos dos conjuntos de teste. Assim a validação cruzada evita a superposição dos conjuntos de teste permitindo a estimativa de um valor de precisão mais real.

Do conjunto de teste O , metade das etiquetas de cada objeto o são selecionadas aleatoriamente e incluídas em Lo , conjunto de etiquetas associadas ao objeto o . A outra metade é incluída em Yo , o gabarito de o . Para a avaliação das estratégias de recomendação em cenários de *cold start*, as etiquetas em Lo são ignoradas. Além disso, foram utilizadas título e descrição como atributos textuais em Fo , para cada objeto o .

O desempenho dos métodos experimentados nas recomendações em cada rodada é avaliada neste trabalho em termos de precisão nas n (com $n = 5$) primeiras posições dos termos recomendados, como feito em (SIGURBJÖRNSSON; Van Zwol, 2008) e (MENEZES et al., 2010). A métrica $P@n$ é dada como segue:

$$P@n(C_o, Y_o) = \frac{|C_{n,o} \cap Y_o|}{\min(n, |Y_o|)}$$

onde, C_o é o conjunto de etiquetas recomendadas para o objeto o , ordenada pelo método avaliado, e Y_o é conjunto de etiquetas consideradas relevantes que formam o gabarito para o objeto o .

Por exemplo, se as 5 primeiras etiquetas retornadas por uma recomendação são {relevante, irrelevante, irrelevante, relevante, relevante} então os valores de $P@1$ a $P@5$ são {1, 1/2, 1/3, 2/4, 3/5}, respectivamente. Assim o desempenho da recomendação é definida como a média $P@n$ sobre todas as recomendações produzidas pelo método avaliado.

3.4 Ambiente de Experimentação

Na Figura 3.3, apresenta-se o ambiente de experimentação construído para a condução dos estudos. O ambiente é composto por um conjunto de *scripts* que estão distribuídos em três módulos: pré-processamento, métodos e avaliação.

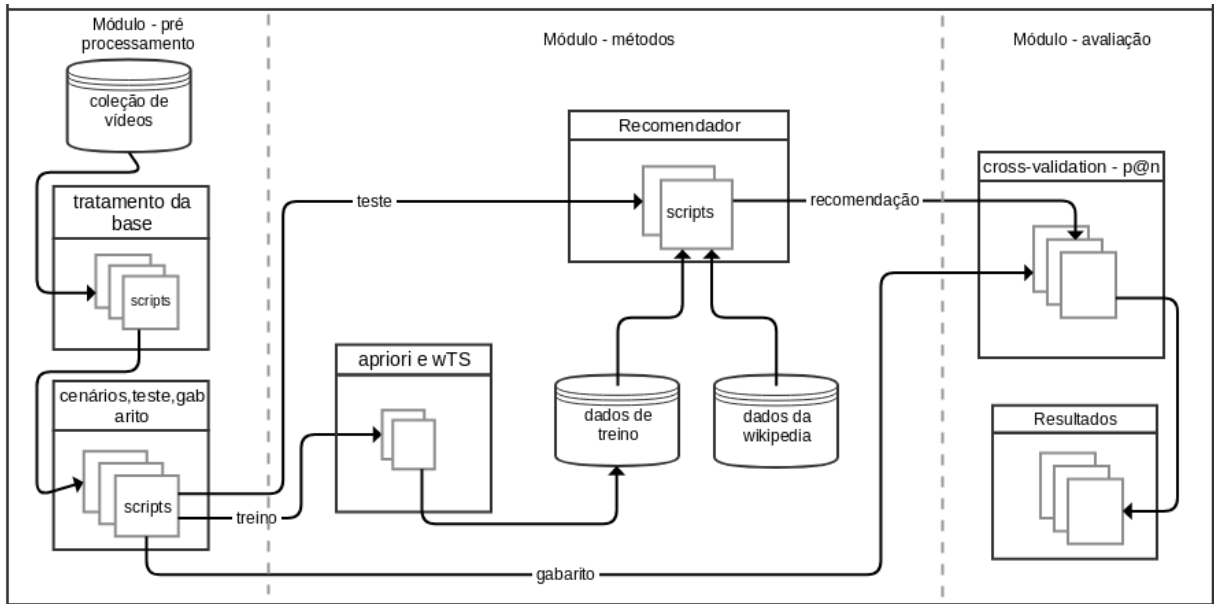


Figura 3.3: Ambiente de experimentação automática.

O módulo de pré-processamento é responsável por realizar o tratamento dos dados oriundos das coleções citadas na Seção 3.2, e pela definição dos cenários, conjuntos de teste, conjuntos de treino e gabarito citados na Seção 3.3. O tratamento realiza a remoção de termos *stopwords*, identificação e a validação de n-gramas. Após essa etapa os dados

passam por *scripts* responsáveis por definir a que cenário um vídeo pertence de acordo com a quantidade de etiquetas que possui, a escolha do vídeo para compor a amostra de treino ou de teste. E por fim, a escolha da etiqueta para formar o gabarito ou etiqueta associada ao objeto.

O módulo métodos é responsável pela implementação dos métodos estudados neste trabalho: i) Sum^+wTS que também engloba o algoritmo apriori; ii) os Grafos de Contexto (GC) e iii) os algoritmos de análise de *links*. O conjunto de treino gerado no módulo pré-processamento é utilizado pelo Apriori e wTS para a geração das regras de associação e valores de wTS dos atributos textuais do conjunto e são armazenados na base de dados para então fornecerem as entradas dos *scripts* que compõem a próxima etapa. A etapa Recomendador é a que produz a recomendação das etiquetas, oriundas da implementação do método base e da estratégia proposta abordadas neste trabalho. Essa etapa, que é executada de forma independente, consome os conjuntos de teste e os dados da Wikipédia e geram as recomendações que serão analisadas pelo módulo de avaliações.

O módulo avaliações recebe as recomendações do módulo métodos e as submete a métrica $P@n$ descrita na Seção 3.3 para avaliar o desempenho dos métodos utilizando o gabarito com as etiquetas consideradas relevantes, sendo os resultados armazenados ao final do processo.

3.5 Resultados numéricos

Nesta seção apresentam-se os resultados dos experimentos realizados para a validação e avaliação da implementação do método Sum^+wTS em diferentes contextos. Considerou-se os resultados em três amostras de dados: Youtube_ufmg, Youtube_Sports, e a Youtube_Sports_rev. Esta última amostra é resultado da revisão da amostra Youtube_Sports.

A Tabela 3.5 mostra os valores escolhidos para a parametrização do método nos quatro subconjuntos de etiquetas, que foram obtidos a partir da média dos resultados da validação cruzada no conjunto de validação, descrito na Seção 3.3, oriundos das amostras. Os parâmetros σ_{min} e θ_{min} referem-se aos valores de confiança e de suporte responsáveis pelo número de regras de associação gerados e, por conseguinte tempo de processamento na recomendação. k_x , k_c , k_r são parâmetros de ajuste, e o parâmetro α pondera os valores obtidos pelas métricas de recomendação.

Assim, os valores apresentados representam a medida necessária para obtenção dos melhores resultados do método visando o equilíbrio entre precisão e tempo de processa-

mento na recomendação.

Tabela 3.1: Parâmetros de análise das estratégias de recomendação.

Parâmetros	Subconjuntos de etiquetas			
	Misto	Menor	Médio	Maior
k_x, k_c, k_r	5			
σ_{min}	1	1	1	1
θ_{min}	0.1	0.1	0.1	0.2
α	0.7	0.8	0.8	0.7

3.5.1 Primeira rodada - amostra Youtube_ufmg

Na Tabela 3.2 mostram-se os resultados de $P@5$ do método Sum^+wTS na amostra de dados Youtube_ufmg. Os resultados apresentados são semelhantes aos mostrados em (BELÉM et al., 2011) na ausência de *cold start*. A diferença no desempenho é cerca de 5% na maioria dos subconjuntos de etiquetas. Observa-se uma tendência de uma melhoria no desempenho do método em subconjuntos homogêneos em relação a soluções com vídeos que possuem quantidades variadas de etiquetas. Por exemplo, o valor de $P@5$ do método no subconjunto de etiquetas *menor* é 7% maior do que o obtido para o subconjunto de etiquetas *misto*.

Na presença de *cold start* o único resultado disponível é o desempenho do método Sum^+wTS no subconjunto de etiquetas *misto* presente em (MARTINS et al., 2013). O desempenho obtido neste estado é similar ao registrado em (MARTINS et al., 2013), com cerca de 6% de diferença. Nesse subconjunto de etiquetas, o método apresenta uma queda de 20% em seu desempenho.

No cenário homogêneo, observou-se um maior impacto no desempenho do Sum^+wTS no subconjunto de etiquetas *menor* com uma queda de 31%, devido a ausência de termos oriundos da coocorrência que proporciona a busca por novos termos além dos que estão presentes nos atributos textuais do objeto, que neste cenário possuem uma quantidade restrita. Nos subconjuntos de etiquetas *médio* e *maior* o impacto no desempenho do método diminui, as quedas são de 19% e 8%, respectivamente. Isto ocorre pois os vídeos apresentam um maior vocabulário.

Tabela 3.2: Desempenho do método Sum^+wTS com a base Youtube_ufmg.

Estado	Subconjuntos de etiquetas			
	Misto	Menor	Médio	Maior
Sem coldstart	0.475	0.510	0.481	0.426
Sem coldstart (BELÉM et al., 2011)	0.502	0.593	0.461	0.431
Coldstart	0.381	0.351	0.393	0.399
Coldstart (MARTINS et al., 2013)	0.359	—	—	—

Assim os resultados obtidos validam a implementação do método Sum^+wTS e permite a sua avaliação utilizando as amostras coletadas para este trabalho.

3.5.2 Segunda rodada - amostra Youtube_Sports

Na Tabela 3.3 mostram-se os resultados de $P@5$ do método Sum^+wTS na amostra Youtube_Sports. O Sum^+wTS obteve desempenho abaixo do apresentado na amostra Youtube_ufmg, tanto na ausência quanto na presença de *cold start* e em todos os subconjuntos de etiquetas. Existe alguns pontos a serem considerados e que justificam a diferença entre resultados envolvendo essas duas coleções:

- 1) O Pré-processamento realizado na amostra Youtube_ufmg como remoção de afixos e plurais. Este processo não foi realizado em nossa base devido a sua utilização com a abordagem dos grafos de contexto que utilizam artigos da Wikipédia.
- 2) As bases são formadas por vídeos de categorias diferentes e são de línguas diferentes. A amostra Youtube_ufmg é composta por vídeos de diversas categorias e se limita a vídeos onde seus atributos textuais são formados por termos da língua inglesa, enquanto que a amostra Youtube_Sports é formada por vídeos da categoria esportes e os termos em sua maioria estão em língua portuguesa.
- 3) Os usuários não atribuem etiquetas que expressam o conteúdo do vídeo postado. Uma análise na base Youtube_Sports mostra que as etiquetas são usadas para melhorar a indexação dos vídeos, uma vez que elas são um dos fatores que influenciam o ranqueamento dos vídeos na busca. Além disso, muitos canais apenas replicam conteúdo, e para que seus vídeos sejam indexados e vistos nas melhores posições nas

pesquisas, acabam definindo etiquetas que promovem o canal e/ou escolhem termos que estão em voga na aplicação. Por exemplo, no contexto esportivo são utilizados jogadores e times famosos. Esta forma de atribuição de etiquetas é denominada de *spamming* e vem prejudicando os métodos de recomendação, por serem dependentes dos dados disponibilizados pelos usuários.

Tabela 3.3: Desempenho do método Sum^+wTS com a amostra Youtube_Sports

Estado	Subconjuntos de etiquetas			
	Misto	Menor	Médio	Maior
sem coldstart	0.353	0.451	0.366	0.308
coldstart	0.337	0.325	0.303	0.383

O fato da amostra Youtube_Sports apresentar tais ruídos e tendo-se extraído o gabarito dessa amostra, podem inferir que os resultados obtidos com o Sum^+wTS não representam sua real eficiência. Visando a redução do impacto desses ruídos, revisou-se a amostra Youtube_Sports validando-se suas etiquetas e uma nova rodada de experimentos foi conduzida.

3.5.3 Terceira rodada - amostra Youtube_Sports revisada

A amostra Youtube_Sports revisada, denominada Youtube_Sports_Rev, possui 3000 vídeos que foram retirados aleatoriamente da amostra Youtube_Sports. A revisão teve como objetivo a remoção de etiquetas que enquadram-se nas características do terceiro fator apresentado na Seção 3.5.2. Também formou-se uma amostra que será denominada de Youtube_Sports_3000 com os mesmos vídeos da amostra Youtube_Sports_Rev, entretanto os metadados desses vídeos são os originais presentes da amostra Youtube_Sports com os ruídos. Desta maneira pode-se avaliar o impacto dos termos *spams* por meio destas duas amostras.

As características citadas realmente impactam no desempenho do método Sum^+wTS . Nos subconjuntos de etiquetas *misto*, *menor* e *maior* nos estados de ausência e presença de *cold start* ocorreu uma melhoria de cerca de 22%-30% nos valores de $P@5$ das recomendações em comparação com a coleção Youtube_Sports_3000, ver Tabelas 3.4 e 3.5.

Tabela 3.4: Desempenho do método Sum^+wTS com amostra Youtube_Sports_3000.

Estado	Subconjuntos de etiquetas			
	<i>Misto</i>	<i>Menor</i>	<i>Médio</i>	<i>Maior</i>
sem coldstart	0.366	0.405	0.463	0.415
coldstart	0.351	0.293	0.285	0.521

Em relação ao comportamento do método Sum^+wTS na amostra Youtube_Sports_Rev, os subconjuntos de etiquetas *misto*, *menor* e *médio* foram impactados pelo *cold start*. O maior impacto foi apresentado no subconjunto *menor* com uma queda de 18% no desempenho, em seguida no subconjunto *médio* com 17% e no subconjunto *misto* com 10%. Em geral, os vídeos que apresentam mais etiquetas possuem um maior vocabulário em seus atributos textuais. Os subconjuntos de etiquetas *menor* e *médio* possuem um vocabulário 41% e 30% menor, respectivamente, do que o subconjunto *maior*, o que justifica a queda do impacto no desempenho do Sum^+wTS na ausência de etiquetas de acordo com os subconjuntos de etiquetas.

Destaca-se o aumento do desempenho do Sum^+wTS no subconjunto de etiquetas *maior*, que diferente dos outros subconjuntos apresenta um aumento de 19% no valor de $P@5$ em *cold start*. O maior vocabulário em seus atributos textuais favorece a heurística wTS , que em *cold start*, torna-se o único método que atua no ranqueamento das recomendações, e a tendência é que esse vocabulário seja suficiente para descrever o vídeo.

Tabela 3.5: Desempenho do método Sum^+wTS com a amostra Youtube_Sports_Rev.

Estado	Subconjuntos de etiquetas			
	<i>Misto</i>	<i>Menor</i>	<i>Médio</i>	<i>Maior</i>
sem coldstart	0.467	0.51	0.443	0.364
coldstart	0.420	0.42	0.369	0.451

3.6 Considerações Finais do Capítulo 3

Neste capítulo foram apresentados o método Sum^+wTS , o arcabouço utilizado nos experimentos com uma abordagem de avaliação automática e os resultados obtidos com experimentos que avaliaram o método. Este arcabouço foi implementado seguindo a metodologia utilizada em (BELÉM et al., 2011), que influenciou na forma como esta

dissertação foi construída.

A coleta dos dados foi um dos processos que mais demandou tempo, já que inicialmente o Youtube não compartilhava as etiquetas atribuídas aos vídeos, obrigando-nos a partir para alternativas que pudessem definir etiquetas para os vídeos por meio de heurísticas. No entanto, o cenário muda e a aplicação passa a compartilhar este dado e então iniciou-se a coleta dos dados dos vídeos. Apesar do Youtube permitir o compartilhamento das etiquetas, ao realizar as coletas percebeu-se que maior parte dos vídeos esportivos não possuem conteúdo em português dificultando a coleta de mais de 150 mil vídeos. Os dados coletados foram utilizados nas avaliações dos desempenhos dos métodos.

A metodologia aborda a caracterização dos cenários, que buscou simular os diferentes contextos que as recomendações podem assumir de acordo com a quantidade de etiquetas que os usuários fornecem a aplicação. Além do procedimento experimental de validação cruzada, a métrica $P@n$, e o ambiente de experimentação utilizados nos experimentos. Sendo assim este arcabouço é responsável pela realização de uma avaliação quantitativa dos métodos avaliados.

Os resultados obtidos com os experimentos avaliaram o método Sum^+wTS em todos cenários e estados definidos. Verificou-se na primeira rodada de experimentos com a amostra Youtube_ufmg que o Sum^+wTS implementado neste trabalho alcançou resultados semelhantes aos encontrados em (MARTINS et al., 2013). Os experimentos realizados com a amostra Youtube_Sports revelaram um desempenho abaixo do esperado, fruto dos ruídos mencionados que impactaram negativamente no desempenho do método.

Com a finalidade de verificar o impacto dos ruídos no desempenho do Sum^+wTS , realizou-se uma revisão da amostra e uma nova rodada de experimento, tanto com a base original quanto com a revisada. Os resultados mostram uma melhoria no desempenho do método nos subconjuntos de dados *misto* e *menor* no estado sem *cold start* e melhoria em todos os cenários em *cold start*, adequando os resultados para um cenário ideal, sem ruídos.

Na coleção formada para este trabalho, o método Sum^+wTS manteve os mesmos resultados em (MARTINS et al., 2013) em relação ao impacto gerado pelo *cold start* na maioria dos cenários, com exceção no subconjunto *maior*. Sem as etiquetas provenientes da coocorrência, a recomendação realizada resume-se a etiquetas candidatas oriundas dos atributos textuais justificando a queda no desempenho do método.

É importante ressaltar que em aplicações colaborativas, nem sempre os usuários preocupam-se em fornecer dados corretos, as coleções estão suscetíveis a termos com

erros de grafia, que não descrevem o conteúdo do objeto, o que torna o *wTS* influente em *cold start* por valorar tais termos candidatos pela presença nos atributos textuais. A presença de termos spammers na amostra Youtube_Sports é um exemplo do impacto negativo nas recomendações, i.e. cerca de 12% no subconjunto de etiquetas *menor* em *cold start*. Os usuários utilizam este meio para tornarem seus vídeos mais visualizados, sendo essa prática cada vez mais presente em aplicações colaborativas.

Capítulo 4

Estratégia de recomendação associativa de etiquetas usando grafos de contexto

Neste Capítulo, apresenta-se uma estratégia de recomendação de etiquetas, supondo-se: i) que os objetos a serem etiquetados são vídeos de curta duração, com temática específica; e ii) que não há etiquetas previamente associadas aos objetos, caracterizando o problema de *cold start*. A especificidade da temática impacta o tamanho do conjunto de etiquetas relevantes, tornando ainda mais sensível o processo de recomendação, e o estado de *cold start* limita as abordagens baseada em coocorrência pois os termos candidatos, usados no processo de expansão, são oriundos de fontes textuais ruidosas associadas ao objeto. O processo de expansão da estratégia proposta emprega o contexto criado pelas relações entre termos presentes na Wikipédia.

Os experimentos realizados capturam o comportamento dos grafos com sua integração com os conjuntos de recomendações vindos do método *Sum⁺wTS* em cenários com amostras de vídeos que possuem um conjunto de etiquetas homogêneo e heterôgeneo. Diferentes algoritmos de análise de *links* são utilizadas para valorar os conjuntos recomendados e avalia-se de que forma os grafos de contexto impactam o ranqueamento final desses conjuntos.

Este capítulo organiza-se da seguinte forma: na Seção 4.1 apresenta-se o método que aplica o conceito de grafo de contexto na busca de conceitos em textos curtos e sua composição. Na Seção 4.2 apresenta-se a estratégia de recomendação de etiquetas com o uso dos grafos de contexto proposta nesta dissertação, bem como sua composição.

Na Seção 4.3 apresenta-se a análise dos resultados numéricos obtidos nos experimentos realizados. E por fim, na Seção 4.4 estão as considerações finais deste capítulo.

4.1 Encontrando conceitos em textos curtos

O compartilhamento de conteúdo na *web*, como fotos, vídeos, e músicas, é cada vez mais difundido e crescente. Além disso, redes sociais, como facebook, twitter, permitem a expressão de sentimentos, pensamentos, opiniões de seus usuários, na forma de *posts*. Encontrar o significado nesse tipo de texto representa uma etapa importante na extração de informações valiosas contida na imensa quantidade de dados fornecidas pelos usuários.

Neste contexto, em (BERLT et al., 2015) apresenta-se um método que utiliza as informações de conectividade presentes na Wikipédia para descobrir os conceitos mais relacionados a textos curtos. Além disso, o método proposto é utilizado como um novo conjunto de funcionalidades em um *framework* de aprendizagem de máquina para aumentar a qualidade dos resultados obtidos.

A suposição é de que quando duas ou mais palavras aparecem juntas em um fragmento de texto curto e essas palavras estão interligadas dentro da Wikipédia, é provável que elas estejam relacionadas com a temática do texto. A conexão dos artigos é modelada usando grafo de contexto, que representa as palavras chaves encontradas no texto como vértices e o relacionamento entre elas como arestas.

O método define quatro tipos de grafo, que juntos representam diferentes aspectos da conectividade dos termos presentes na Wikipédia. Além disso, o método usa algoritmos de análise de *link* para ordenar os vértices de acordo com sua relevância no grafo. Nas seções seguintes tais grafos são apresentados, bem como os algoritmos de análise de *links* empregados.

4.1.1 Full Local Context Graph - *flcg*

O primeiro grafo, chamado *Full Local Context Graph* (*flcg*), representa a conectividade de todos os termos presentes no conjunto inicial de termos candidatos na Wikipédia. Assim, o grafo *flcg*, resultante do conjunto inicial de termos candidatos T , é definido como:

$$G_{flcg}(T) = \langle V_{flcg}, E_{flcg} \rangle$$

O conjunto de vértices V_{flcg} é obtido da seguinte forma: os n-gramas presentes no conjunto inicial de termos são extraídos e compõem o conjunto inicial de vértices candidatos. Desse conjunto, os n-gramas que possuem pontuações entre duas das suas palavras são removidas. Em seguida, os n-gramas que não combinam com qualquer título de artigo da Wikipédia na forma normalizada, também são removidos. Então, o conjunto resultante representa o conjunto de vértices V_{flcg} .

Normaliza-se os títulos de artigos da Wikipédia evitando a ambiguidade dos termos. As pontuações foram retiradas e todos os caracteres dos termos que compõem o título do artigo foram formatados para a forma minúscula.

O conjunto de arestas E_{flcg} é obtido da seguinte forma: cada vértice v presente em V_{flcg} corresponde a um artigo no Wikipédia $w(v)$, desta maneira, só existirá aresta $e = (v, u)$, do vértice v para o vértice u , se e somente se, existir uma ligação entre $w(v)$ e $w(u)$ na Wikipédia, ou seja, se existir um *link* entre os artigos representados pelos vértices v e u .

A Figura 4.1 mostra o grafo $flcg$ resultante do conjunto inicial de termos candidatos obtido do seguinte título e descrição de um vídeo:

Título: Pelé, o futebolista do século pela fifa

Descrição: Carreira do camisa 10 da Seleção Brasileira, de rei do futebol a compositor de música

O conjunto inicial contém onze n-gramas, gerando onze vértices distintos.

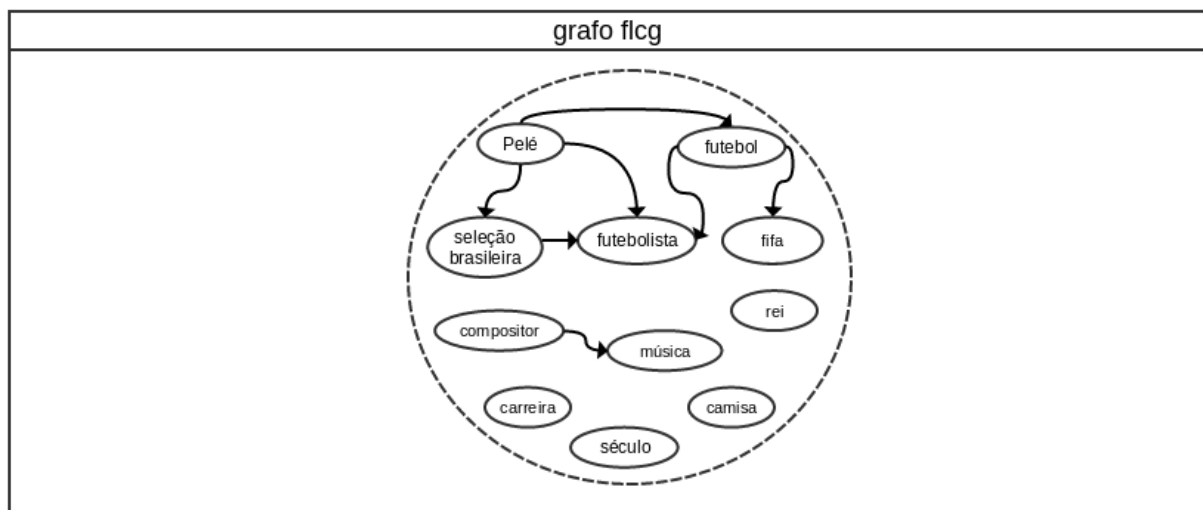


Figura 4.1: *Full Local Context Graph - flcg*

4.1.2 Degree Local Context Graph - *dlcg*

O segundo grafo, chamado de *Degree Local Context Graph*, é uma variação do grafo *flcg*. A diferença é que todos os vértices com grau zero são removidos do conjunto de vértices V_{flcg} , devido alguns termos presentes no título e na descrição não terem relação com o assunto principal do contexto na qual o vídeo está inserido. Em outras palavras, tais termos são conceitos irrelevantes que não se conectam com nenhum outro conceito no grafo. Assim o *dlcg* é definido como:

$$G_{dlcg}(T) = \{\langle V_{dlcg} \rangle, \langle E_{dlcg} \rangle\}$$

onde $E_{dlcg} = E_{flcg}$. Nesse grafo, o conjunto de vértices V_{dlcg} é dado por $\{ u \in V_{dlcg} \mid (u, v) \vee (v, u) \in E_{dlcg} \}$. A Figura 4.2 mostra o grafo *dlcg* resultante após a remoção dos vértices que não possuem conexão.

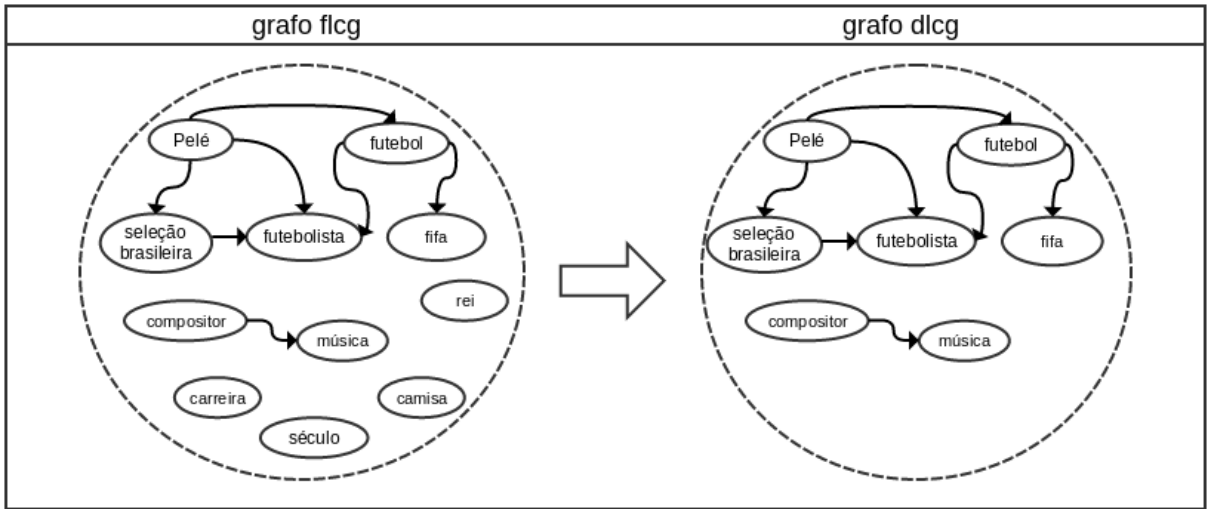


Figura 4.2: *Degree Local Context Graph - dlcg*

4.1.3 Connected Local Context Graph - *clcg*

O *Connected Local Context Graph* (*clcg*) também é uma variação do grafo *flcg*. Esse terceiro grafo identifica o maior conjunto de conceitos relacionados que são gerados no grafo *flcg*.

A construção desse grafo é motivada pelo fato que, em alguns casos, dois ou mais conjuntos de conceitos estão relacionados uns com os outros no grafo *dlcg*, mas não tem nenhuma relação com o assunto principal. A Figura 4.3 mostra o grafo *clcg* resultante com o maior conjunto de vértices do conjunto inicial de termos T .

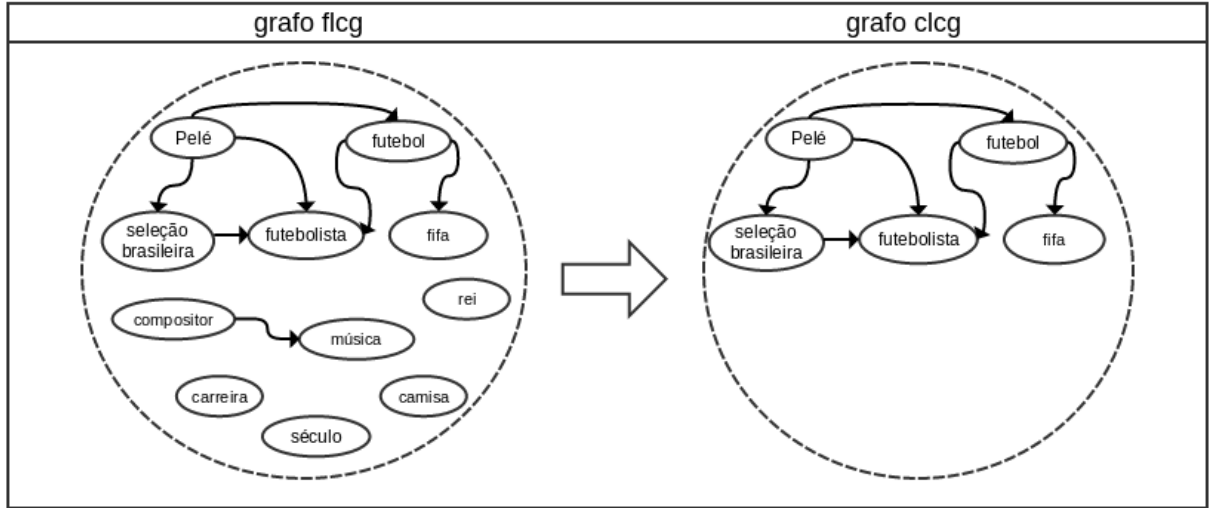


Figura 4.3: *Connected Local Context Graph - clcg*

4.1.4 Strong Local Context Graph - *slcg*

O grafo *Strong Local Context Graph (slcg)* trata o problema da falta de transitividade entre os conceitos, que pode incluir conceitos irrelevantes para a busca do significado do título e descrição de um dado vídeo. Por exemplo, existe relação entre os conceitos a e b , assim como existe relação entre os conceitos b e c , entretanto não existe relação entre os conceitos a e c . O *slcg* é definido como segue:

$$G_{slcg}(T) = \langle V_{slcg}, E_{slcg} \rangle.$$

onde V_{slcg} é o conjunto maximal de vértices, que satisfaz a seguinte restrição: todo par de vértices $u, v \in V_{slcg}$, e ao menos umas das arestas (u, v) ou $(v, u) \in E_{flcg}$. A Figura 4.4 mostra o grafo *slcg* resultante com os termos mais relacionados do conjunto inicial de termos T .

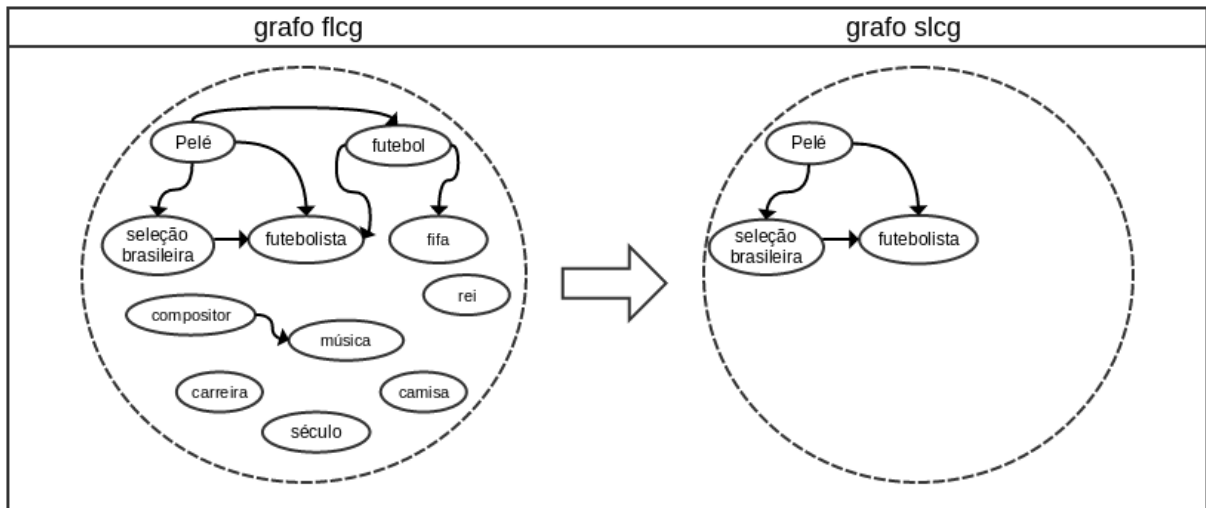


Figura 4.4: *Strong Local Context Graph - slcg*

4.1.5 Expanded Contexts Graphs

Uma importante característica do Wikipédia é a presença de páginas que redirecionam os usuários a outras páginas que contem a informação que os usuários estão procurando. Denominadas como redirecionamentos, essas páginas são úteis pois ajudam a evitar a ambiguidade de certos conceitos, que podem ser descritos de diversas formas.

Ao considerar as páginas de redirecionamentos, os grafos de contexto desenvolvidos são extendidos, pela adição deste tipo de página como vértice do grafo, sendo chamados como Expanded Contexts Graphs. Essa adição é responsável por incorporar de forma sintática e semântica o conceito correto com seu termo (etiqueta candidata a recomendação), sendo aqui denominado de etiquetas formais. A Figura 4.5 mostra o grafo *xflcg* resultante do grafo *flcg*:

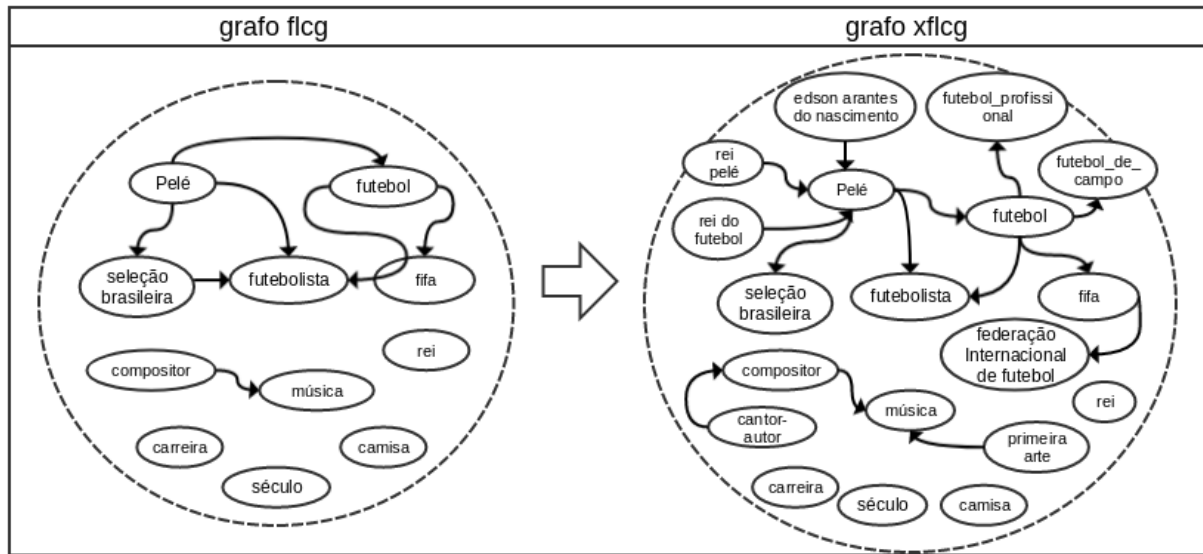


Figura 4.5: Expanded Full Local Contexts Graphs - *xflcg*

Os grafos estendidos são denominados da seguinte forma: *xflcg*, *xdlcg*, *xclcg* e *xslcg*.

4.1.6 Algoritmos de análise de *links*

Após a construção dos grafos de contexto por meio do conjunto inicial de termos candidatos, utilizam-se algoritmos de análise de *links* para classificar os vértices dos grafos. Assim retorna-se uma lista ordenada dos conceitos relacionados ao conjunto inicial de termos dado. Esses algoritmos analisam as ligações entre os vértices para predizer a ordenação, posicionando os conceitos mais relacionados no início da lista. Os algoritmos usados são:

- Indegree e Outdegree - algoritmo que pontua vértices de acordo com o seu grau, sendo que o Indegree pontua os vértices de acordo com o número de arestas que aponta para ele, e o Outdegree pontua os vértices de acordo com o número de arestas que apontam para outros vértices do grafo. Assim como em (BERLT et al., 2015), serão consideradas apenas as arestas presentes no grafo de contexto para calcular o valor de cada vértice.
- HITS - Neste algoritmo, os vértices recebem duas pontuações, seguindo os conceitos de *hub* e *authority*. O *hub* é um vértice com muitos *links* de entrada e o *authority* um vértice que possui muitos *links* de saída. Desta forma, considerado-se um bom *hub*, um vértice que aponta para bons vértices do tipo *authority*. Para ser considerado

um bom vértice do tipo *authority*, um vértice deve apontar para bons vértices do tipo *hub*. Assim, as duas pontuações são consideradas, para realizar o ranqueamento dos vértices do grafo.

- PageRank - Neste algoritmo, o cálculo da pontuação de cada vértice é realizado de maneira recursiva e leva em conta, os valores dos vértices que apontam para ele. Assim, o valor de PageRank de cada vértice se propaga para os demais vértices que aponta. É uma técnica robusta e que define a pontuação de um vértice em um único valor.

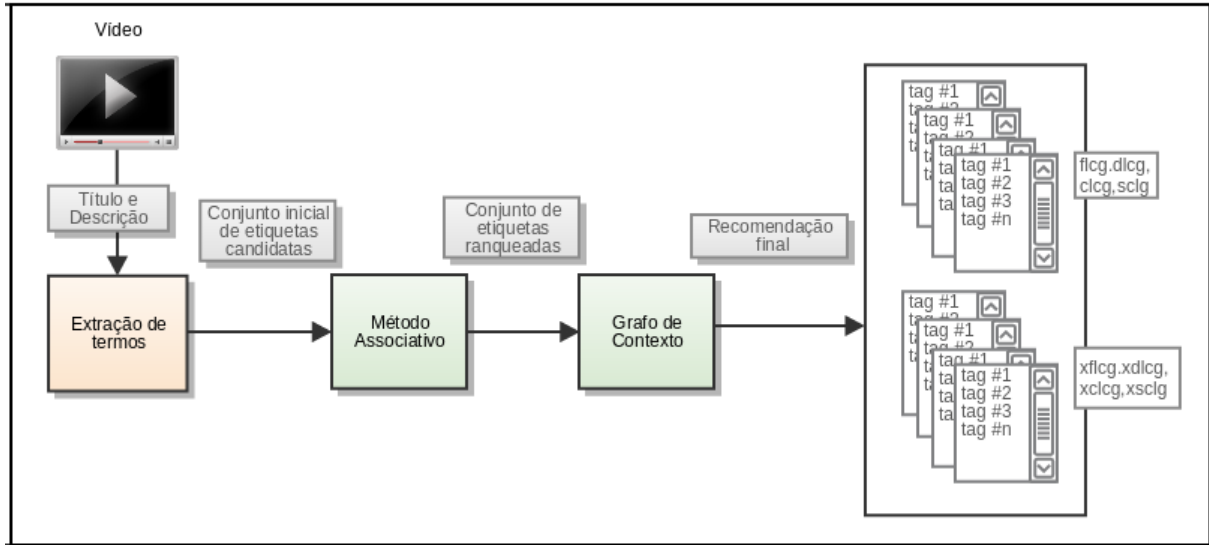
4.2 Estratégia de recomendação associativa de etiquetas com grafos de contexto

Métodos associativos abordam o problema de recomendação de etiquetas explorando padrões de coocorrência, a partir de etiquetas previamente associadas ao objeto. Esse tipo de estratégia vem constatemente obtendo resultados estado da arte. Em cenário de *cold start*, devido o emprego da coocorrência, essas estratégias são pouco utilizadas.

Os trabalhos que abordam o problema de *cold start* utilizam atributos textuais do objeto, métricas de relevância de etiquetas e etiquetas associadas a objetos similares. Nos últimos anos, o conceito de *relevance feedback* vem sendo aplicado para melhorar os resultados desses trabalhos, tirando proveito do engajamento e do *feedback* dos usuários. Entretanto, nem sempre esse engajamento pode ser identificado nos usuários o que dificulta o uso dos métodos baseados nesse conceito em aplicações reais.

Assim, a estratégia de recomendação de etiquetas baseada em grafo de contexto, denominada Sum^+wTS^+GC , tem como objetivo a recomendação de etiquetas relevantes ao mesmo tempo que auxilia os usuários a escolherem etiquetas que sejam mais conceituais tirando proveito das informações presentes em uma enciclopédia colaborativa. Ou seja, busca-se a redução do impacto de problemas relacionados a qualidade das etiquetas atribuídas aos vídeos. Além de analisar o comportamento dos usuários a diferentes modos de recomendação permitidos pelo emprego dos grafos de contexto.

Na Figura 4.6 mostra-se o fluxo da estratégia proposta para a realização da recomendação das etiquetas. A estratégia é composta por três etapas: extração dos termos, método associativo e grafo de contexto. Ao final do fluxo, as etiquetas recomendadas são analisadas quanto a sua relevância. A seguir, as etapas do fluxo de recomendação são apresentadas em detalhes.

Figura 4.6: Esquema da estratégia Sum^+wTS^+GC

4.2.1 Extração dos termos

O conjunto inicial de termos candidatos é composto por termos extraídos de duas fontes textuais associados ao vídeo alvo da recomendação, o título e a descrição. Essas fontes são tratadas removendo-se termos com alta frequência e pouca importância como artigos, preposições, e pronomes. Em seguida os termos resultantes são submetidos a uma função que gera até trigramas, a fim de obter termos compostos comuns no vocabulário esportivo, i.e. gol de placa.

Um glossário de bigramas e trigramas foi montado para validar os n-gramas gerados pela função. Esses n-gramas que compoem o glossário são oriundos da base de títulos de todos os artigos da Wikipédia criados por brasileiros (domínio pt). Esta base de títulos possui cerca de um milhão e oitocentos mil termos, entretanto, cerca de 600 mil termos foram incluídos no glossário.

Após esse processo, os termos resultantes formam o conjunto inicial de termos candidatos, que serão modelados no próximo passo da estratégia.

4.2.2 Método associativo

O segundo passo da estratégia proposta é a utilização da lista de termos candidatos resultante do passo anterior. Esses termos serão a entrada para o Sum^+wTS , o método associativo estado da arte descrito no Capítulo 3. O uso do Sum^+wTS visa uma primeira

análise da relevância do conjunto de termos candidatos. Neste ponto, a relevância é obtida de acordo com a frequência dos termos na base de dados, especificamente das etiquetas previamente associadas aos vídeos, associado ao valor dado por uma heurística que mede a importância do termo por meio da presença nos atributos textuais. O conjunto de etiquetas resultante é composto por termos ranqueados de acordo com as medidas de relevância citadas. A presença de *cold start* irá definir o valor obtido por cada termo.

4.2.3 Grafo de contexto

O terceiro passo da estratégia proposta consiste no uso do método de busca de significado em textos curtos descrito na Seção 4.1, usando a descoberta de conceitos relacionados de um conjunto de termos candidatos a recomendação oriundos da etapa anterior.

Os grafos de contexto reorganizam o conjunto de termos candidatos a recomendação levando em consideração a conexão destes termos na Wikipédia. Esta reorganização ocorre através do refinamento do conjunto recomendado, que neste trabalho é realizado pela análise da presença do termo na Wikipédia, a medição de sua relevância através de suas conexões e pela adição de etiquetas formais ao conjunto inicial de recomendação.

Um conjunto de recomendação vindo da etapa dos métodos associativos produz nove conjuntos de recomendação nesta etapa, cada conjunto é formado por uma ordenação dos termos diferente, que depende de qual aspecto das conexões dos termos na Wikipédia foi analisado por cada tipo de grafo, ver na Seção 4.1.

A ordenação final dos termos candidatos a recomendação considera a ponderação dos valores recebidos de cada termo candidato c pelo método Sum^+wTS e pelo algoritmos de análise de *links* por meio dos grafos de contexto construídos. Dessa forma, esta etapa é responsável pela busca da qualidade conceitual dos termos candidatos a descritores do objeto alvo da recomendação. A Equação 4.1 apresenta a estratégia Sum^+wTS^+GC .

$$Sum^+wTS^+GC(c,o,k_x,k_c,k_r,\alpha) = \alpha Sum^+wTS(c,o,k_x,k_c,k_r) + (1 - \alpha)GC(c,o) \quad (4.1)$$

onde c é a etiqueta candidata a recomendação para o objeto o , k_x , k_c , k_r são parâmetros de ajuste. O parâmetro α pondera os valores obtidos de cada etiqueta pelo Sum^+wTS e pelos algoritmos de análise de *links* em cada grafo de contexto utilizado. Os valores escolhidos para a parametrização do Sum^+wTS^+GC é o mesmo utilizado pelo método Sum^+wTS descrito no Capítulo 3.

4.3 Resultados Numéricos

Nesta seção apresentam-se os resultados dos experimentos realizados para avaliar a estratégia de recomendação proposta. Considerou-se os resultados em duas situações distintas, a primeira refere-se a uma coleção de dados com os dados originais dos vídeos, contendo os ruídos, i.e. termos populares que não estão relacionados ao conteúdo do vídeo, mencionados na Seção 3.5.2, e a segunda refere-se a mesma coleção, entretanto realizou-se uma revisão no seu conteúdo com a finalidade de reduzir o ruído presente na amostra anterior.

4.3.1 Primeira rodada - amostra Youtube_Sports

Essa primeira rodada de experimentos é caracterizada pelo uso de uma coleção de vídeos que apresenta ruído nos seus atributos textuais, título e descrição, e pela ausência e presença do *cold start*. As tabelas 4.1 e 4.2 mostram os resultados obtidos para os cenários gerados pela ausência e ocorrência do *cold start*.

Os resultados apresentados para a estratégia Sum^+wTS^+GC são aqueles obtidos quando o algoritmo de análise de *links* pagerank foi usado, tendo sido este o que apresentou o melhor desempenho entre os vários algoritmos de análise de *links* considerados.

As características da coleção, discutidas na Seção 3.5.2, impactaram mais fortemente a estratégia proposta do que o método base. Em muitos casos, o termo ruído, presente nos atributos textuais, possui grafia distinta ou não possui entradas na Wikipédia. Essas características limitam a atribuição de valor aos termos da coocorrência e dos atributos textuais, o que justifica o impacto observado.

Tabela 4.1: Desempenho do método Sum^+wTS^+GC com a amostra Youtube_Sports no estado sem *cold start*

		Subconjunto de etiquetas			
método proposto		Misto	Menor	Médio	Maior
Sum^+wTS^+GC	<i>flcg</i>	0.191	0.283	0.230	0.185
	<i>dlcg</i>	0.153	0.245	0.193	0.152
	<i>clcg</i>	0.159	0.254	0.213	0.156
	<i>slcg</i>	0.224	0.325	0.302	0.143
	<i>xflcg</i>	0.190	0.270	0.205	0.170
	<i>xdlcg</i>	0.173	0.255	0.183	0.159
	<i>xclcg</i>	0.156	0.256	0.196	0.144
	<i>xslcg</i>	0.222	0.324	0.295	0.140
método base					
Sum^+wTS		0.353	0.451	0.366	0.308

Tabela 4.2: Desempenho do Sum^+wTS^+GC com a amostra Youtube_Sports no estado de *cold start*

		Subconjunto de etiquetas			
método proposto		Misto	Menor	Médio	Maior
Sum^+wTS^+GC	<i>flcg</i>	0.262	0.280	0.278	0.228
	<i>dlcg</i>	0.250	0.272	0.267	0.211
	<i>clcg</i>	0.267	0.283	0.282	0.237
	<i>slcg</i>	0.327	0.312	0.319	0.350
	<i>xflcg</i>	0.250	0.285	0.260	0.205
	<i>xdlcg</i>	0.245	0.285	0.254	0.196
	<i>xclcg</i>	0.271	0.300	0.285	0.230
	<i>xslcg</i>	0.337	0.331	0.327	0.355
método base					
Sum^+wTS		0.337	0.325	0.303	0.383

4.3.2 Segunda Rodada - amostra Youtube_Sports revisada

Nesta segunda rodada de experimentos foram utilizadas as amostras de dados Youtube_Sports_Rev e Youtube_Sports_3000, ver Seção 3.5.3. Os resultados apresentados são o ponto de partida para a análise dos desempenho da estratégia Sum^+wTS^+GC . Pri-

meio analisa-se na ausência de *cold start*, o comportamento da estratégia com o uso dos diferentes grafos de contexto na amostra Youtube_Sports_3000, seguida do impacto do ruído dos termos na referida amostra, e da análise do comportamento da estratégia utilizando diferentes grafos de contexto na amostra Youtube_Sports_Rev. Depois realiza-se a mesma análise da estratégia nas amostras, porém na ausência de etiquetas nos vídeos caracterizando o estado de *cold start*.

A análise dos resultados obtidos do Sum^+wTS^+GC demanda que se considere a influência dos algoritmos de análise de *links* na forma que pretende-se realizar a recomendação, informal ou formal, e da influência do tipo de grafo de contexto utilizado no comportamento dos termos, no conjunto candidato a recomendação gerado pela estratégia.

Em relação aos algoritmos de análise de *links*, destaca-se que a influência na recomendação é maior quando o Sum^+wTS^+GC usa os grafos de contexto expandidos, já que ocorre a adição das etiquetas formais. Nas versões não expandidas dos grafos as ligações existentes resumem-se aos termos oriundos dos vídeos, que nem sempre apresentam ligações de seus conceitos na Wikipedia, sendo mais observada nas versões expandidas.

Os algoritmos indegree e hub têm como característica a análise da importância de um termo através de outros termos que apontam para ele. Essa característica favorece as etiquetas formais, já que as páginas de redirecionamentos apontam para esses termos, tornando-os mais valorizadas dentre os outros no conjunto de recomendação, principalmente nos grafos expandidos. A importância de um termo nos algoritmos outdegree e authority é dada pelas ligações que saem do termo, ou seja, que apontam para outros termos, no caso dos grafos expandidos, os termos que apontam para as etiquetas formais são os mais valorizados. No pagerank, o valor dado ao termo ocorre através da análise do valor de pagerank dos termos que apontam para ele de forma recursiva, sendo um valor de nível mais global, dessa maneira existe uma distribuição de valores por todos os nós (termos) dos grafos.

Os grafos de contexto podem ser classificados em dois tipos: abrangente e limitador. Os grafos abrangentes são aqueles que possuem uma maior quantidade de termos no grafo, como o *flcg*, *dlcg*, *clcg* e suas extensões, pois consideram desde o maior número de termos conectados (*clcg*) no grafo até todos os termos que possuem artigo na Wikipedia (*flcg*), assim conseguem valorar uma maior quantidade de termos do conjunto recomendado. Enquanto que os grafos limitadores, como o *slcg* e sua extensão *xslcg*, causam menos impacto da valorização dos termos recomendados, pois restringem os grafos aos termos que estão interligados entre si.

Para facilitar a leitura e deixar a descrição mais direta, os resultados apresentados pelo Sum^+wTS^+GC envolvendo os grafos de contexto e os algoritmos de análise de *links*, são referenciados pela composição do nome do grafo mais a abreviação do nome dos algoritmos de análise de *links*. Por exemplo, os resultados associados ao grafo *flcg* que utiliza o algoritmo indegree, são denominados *flcg+ind*. Os algoritmos outdegree, pagerank, hub e authority utilizam as seguintes abreviações: out, pr, hub, e auth, respectivamente.

Nas avaliações que seguem verificou-se o comportamento do Sum^+wTS^+GC utilizando os diferentes grafos de contexto no estado em que as etiquetas estão presentes nos vídeos. Nesse estado, o conjunto final de recomendações pode sofrer influência dos valores relacionados a coocorrência, da presença nos atributos textuais e dos grafos de contexto na valorização dos termos. Na Tabela 4.5 os resultados obtidos da amostra Youtube_Sports_3000 demonstram que o Sum^+wTS^+GC pode auxiliar na valorização dos termos oriundos da coocorrência, entretanto termos ruidosos ganham destaque no ranqueamento devido ao wTS , o que pode-se inferir que os resultados não representam seu real desempenho.

Os resultados obtidos pelo Sum^+wTS^+GC mostram que em todos os subconjuntos de etiquetas, os grafos que apresentam maior impacto no desempenho da estratégia são os grafos *flcg*, *dlcg*, *clcg* e suas versões extendidas, pois são os que mais valoram termos no conjunto candidato a recomendação. Além disso, com a presença dos termos oriundos da coocorrência e dos ruídos, a disposição dos termos valorados pelos grafos fica mais restrito a presença de termos informais que possuem artigo na Wikipédia e de etiquetas formais que já estavam presentes nos atributos textuais, ver Tabela 4.3.

Tabela 4.3: Impacto no desempenho da estratégia Sum^+wTS^+GC pelo uso dos grafos abrangentes.

Subconjunto de etiquetas	Sum^+wTS^+GC	Desempenho
<i>menor</i>	<i>xdlcg+pr</i> , <i>dlcg+pr</i> , <i>xflcg+pr</i>	0.29, 0.31, 0.31
<i>misto</i>	<i>xclcg+out</i> , <i>dlcg+out</i> , <i>clcg+out</i>	0.19, 0.20, 0.20
<i>médio</i>	<i>dlcg+out</i> , <i>clcg+out</i> , <i>xclcg+out</i>	0.30, 0.31, 0.30
<i>maior</i>	<i>xclcg+pr</i> , <i>xdlcg+pr</i> , <i>dlcg+pr</i>	0.24, 0.25, 0.26

O menor impacto no desempenho da estratégia Sum^+wTS^+GC pelo uso dos grafos *slcg* e *xslcg* deve-se a característica limitadora deste tipo de grafo, já que não possuem mais de um termo formal no grafo como um grafo abrangente. Além disso, observa-se nas amostras que em muitos casos nos grafos *flcg* (grafo do qual origina-se os grafos limitadores) pode haver dois ou mais subgrafos de etiquetas que possuem termos conectados entre si, porém o

grafo *slcg* e *xslcg* apenas é formado e valorado por um desses subgrafos, e nem sempre este subgrafo escolhido possui termos relevantes. Desta maneira a estratégia Sum^+wTS^+GC que utiliza este tipo de grafo tem uma menor atuação no ranqueamento do conjunto de recomendação, ver Tabela 4.4.

Tabela 4.4: Impacto no desempenho da estratégia Sum^+wTS^+GC pelo uso dos grafos limitadores.

Subconjunto de etiquetas	Sum^+wTS^+GC	Desempenho
<i>menor</i>	<i>slcg+ind</i> , e <i>xslcg+ind</i>	0.39 e 0.40
<i>misto</i>	<i>slcg+ind</i> , e <i>xslcg+ind</i>	0.27 e 0.27
<i>médio</i>	<i>slcg+out</i> , e <i>xslcg+hub</i>	0.40 e 0.39
<i>maior</i>	<i>slcg+out</i> , e <i>xslcg+out</i>	0.31 e 0.30

Tabela 4.5: Desempenho do método Sum^+wTS^+GC com a amostra Youtube_Sports_3000 em estado sem *cold start*

Subconjunto Menor de etiquetas					
Método proposto:	Algoritmos de análise de <i>links</i>				
Sum^+wTS^+GC	outdegree (out)	indegree (ind)	pagerank (pr)	hub (hub)	authority (auth)
<i>flcg</i>	0,363	0,371	0,338	0,351	0,366
<i>dlcg</i>	0,361	0,363	0,313	0,351	0,366
<i>clcg</i>	0,371	0,373	0,331	0,351	0,366
<i>slcg</i>	0,386	0,393	0,361	0,381	0,388
<i>xflcg</i>	0,360	0,338	0,313	0,345	0,355
<i>xdlcg</i>	0,352	0,333	0,287	0,345	0,355
<i>xclcg</i>	0,370	0,375	0,312	0,352	0,367
<i>xslcg</i>	0,393	0,403	0,366	0,388	0,398
método base					
Sum^+wTS	0,405				
Subconjunto Misto de etiquetas					
Método proposto:	Algoritmos de análise de <i>links</i>				
Sum^+wTS^+GC	outdegree (out)	indegree (ind)	pagerank (pr)	hub (hub)	authority (auth)
<i>flcg</i>	0,229	0,229	0,215	0,220	0,246
<i>dlcg</i>	0,202	0,218	0,202	0,220	0,246
<i>clcg</i>	0,203	0,220	0,210	0,220	0,244
<i>slcg</i>	0,259	0,269	0,260	0,257	0,269
<i>xflcg</i>	0,230	0,238	0,216	0,222	0,247
<i>xdlcg</i>	0,208	0,224	0,203	0,222	0,247
<i>xclcg</i>	0,197	0,218	0,205	0,215	0,238
<i>xslcg</i>	0,259	0,272	0,263	0,256	0,272

método base					
Sum^+_{wTS}	0,366				
Subconjunto Médio de etiquetas					
Método proposto:	Algoritmos de análise de <i>links</i>				
Sum^+_{wTS+GC}	outdegree (out)	indegree (ind)	pagerank (pr)	hub (hub)	authority (auth)
<i>flcg</i>	0,358	0,363	0,352	0,345	0,347
<i>dlcg</i>	0,296	0,339	0,337	0,345	0,347
<i>clcg</i>	0,310	0,339	0,334	0,347	0,349
<i>slcg</i>	0,399	0,388	0,373	0,403	0,389
<i>xflcg</i>	0,372	0,361	0,316	0,334	0,334
<i>xdlcg</i>	0,337	0,316	0,311	0,334	0,334
<i>xclcg</i>	0,303	0,307	0,312	0,326	0,335
<i>xslcg</i>	0,384	0,387	0,371	0,389	0,388
método base					
Sum^+_{wTS}	0,463				
Subconjunto Maior de etiquetas					
Método proposto:	Algoritmos de análise de <i>links</i>				
Sum^+_{wTS+GC}	outdegree (out)	indegree (ind)	pagerank (pr)	hub (hub)	authority (auth)
<i>flcg</i>	0,296	0,290	0,277	0,276	0,273
<i>dlcg</i>	0,271	0,265	0,259	0,276	0,273
<i>clcg</i>	0,258	0,267	0,268	0,279	0,276
<i>slcg</i>	0,314	0,304	0,305	0,313	0,304
<i>xflcg</i>	0,303	0,286	0,261	0,273	0,271
<i>xdlcg</i>	0,279	0,279	0,254	0,273	0,272
<i>xclcg</i>	0,258	0,276	0,246	0,275	0,277
<i>xslcg</i>	0,306	0,307	0,291	0,306	0,309
método base					
Sum^+_{wTS}	0,415				

Assim como no experimento com o método Sum^+wTS , ver Capítulo 3, o Sum^+wTS^+GC foi experimentado utilizando a mesma amostra revisada. Os resultados obtidos mostram uma melhoria dos desempenhos da estratégia nos subconjuntos *menor*, *misto* e *médio* presentes na amostra Youtube_Sports_Rev em relação a amostra Youtube_Sports_3000 na ausência do *cold start*. Cada grafo possui características diferentes entre si, e existe uma variedade na melhoria do desempenho, ver Tabelas 4.5 e 4.6.

A análise das tabelas 4.5 e 4.6 mostra que o Sum^+wTS^+GC utilizando grafos limitadores apresenta um maior aumento no desempenho em relação aos grafos abrangentes na maioria dos subconjuntos de etiquetas. Apesar desse tipo de grafo interferir menos no conjunto de recomendação, a remoção dos termos ruidosos permite uma melhor posição

dos termos que são valorados tanto pelos grafos quanto pela coocorrência.

Houve um aumento no desempenho do Sum^+wTS^+GC de 18%-27% no subconjunto *menor*, o $xdlclg+pr$ apresentou um aumento de 18% no desempenho, o menor dentre os grafos avaliados, enquanto que o maior aumento obtido foi de 27% utilizando o $xslclg+pr$. No subconjunto *médio*, as combinações $dlclg+out$, $clclg+out$, $xclclg+out$ e $xclclg+ind$ apresentaram um aumento de cerca de 14%-15% no desempenho. No subconjunto *misto*, a estratégia obteve aumento de 10% a 33%. As combinações $slclg+out$ e $xslclg+out$ apresentam um aumento entre 33%-35% no desempenho enquanto que o menor aumento é obtido pelos $xflclg+out$ e $xdlclg+ind$ com cerca de 10%. No subconjunto *maior*, a estratégia apresentou uma redução no desempenho de 39%-40% quando as combinações $xflclg+out$ e o $flclg+out$ eram utilizadas. Isso ocorreu porque esse subconjunto de dados apresenta uma maior quantidade de termos nos atributos textuais dos vídeos, quando remove-se termos ruídos que são favorecidos pelo wTS , os termos valorados pelos grafos ganham mais relevância no conjunto de recomendação.

Tabela 4.6: Desempenho do método Sum^+wTS^+GC com a amostra Youtube_Sports_Rev em estado sem *cold start*

Subconjunto menor de etiquetas					
Método proposto:	Algoritmos de análise de <i>links</i>				
Sum^+wTS^+GC	outdegree (out)	indegree (ind)	pagerank (pr)	hub (hub)	authority (auth)
<i>flcg</i>	0,46	0,447	0,44	0,457	0,462
<i>dlcg</i>	0,44	0,425	0,4	0,457	0,462
<i>clcg</i>	0,455	0,462	0,417	0,46	0,472
<i>slcg</i>	0,502	0,50	0,497	0,5	0,497
<i>xflcg</i>	0,465	0,457	0,385	0,47	0,482
<i>xdlcg</i>	0,442	0,432	0,35	0,47	0,482
<i>xclcg</i>	0,46	0,462	0,42	0,47	0,485
<i>xslcg</i>	0,515	0,515	0,502	0,512	0,512
método base					
Sum^+wTS	0,51				
Subconjunto Misto de etiquetas					
Método proposto:	Algoritmos de análise de <i>links</i>				
Sum^+wTS^+GC	outdegree (out)	indegree (ind)	pagerank (pr)	hub (hub)	authority (auth)
<i>flcg</i>	0,268	0,308	0,320	0,284	0,332
<i>dlcg</i>	0,256	0,304	0,266	0,284	0,332
<i>clcg</i>	0,257	0,314	0,274	0,283	0,333
<i>slcg</i>	0,401	0,404	0,360	0,397	0,395
<i>xflcg</i>	0,253	0,263	0,253	0,257	0,301
<i>xdlcg</i>	0,243	0,248	0,226	0,257	0,301

<i>xclcg</i>	0,242	0,282	0,264	0,263	0,311
<i>xslcg</i>	0,387	0,387	0,347	0,383	0,378
método base					
<i>Sum</i> ⁺ <i>wTS</i>	0,467				
Subconjunto Médio de etiquetas					
Método proposto:	Algoritmos de análise de <i>links</i>				
<i>Sum</i> ⁺ <i>wTS</i> ⁺ <i>GC</i>	outdegree (out)	indegree (ind)	pagerank (pr)	hub (hub)	authority (auth)
<i>flcg</i>	0,357	0,365	0,361	0,353	0,367
<i>dlcg</i>	0,347	0,357	0,312	0,353	0,367
<i>clcg</i>	0,362	0,372	0,350	0,357	0,373
<i>slcg</i>	0,412	0,413	0,386	0,410	0,413
<i>xflcg</i>	0,331	0,295	0,294	0,345	0,328
<i>xdlcg</i>	0,312	0,267	0,246	0,345	0,328
<i>xclcg</i>	0,350	0,362	0,335	0,357	0,366
<i>xslcg</i>	0,405	0,409	0,371	0,403	0,408
método base					
<i>Sum</i> ⁺ <i>wTS</i>	0,443				
Subconjunto Maior de etiquetas					
Método proposto:	Algoritmos de análise de <i>links</i>				
<i>Sum</i> ⁺ <i>wTS</i> ⁺ <i>GC</i>	outdegree (out)	indegree (ind)	pagerank (pr)	hub (hub)	authority (auth)
<i>flcg</i>	0,177	0,191	0,179	0,168	0,178
<i>dlcg</i>	0,154	0,164	0,162	0,168	0,178
<i>clcg</i>	0,157	0,171	0,165	0,17	0,18
<i>slcg</i>	0,229	0,24	0,226	0,223	0,243
<i>xflcg</i>	0,185	0,199	0,181	0,167	0,176
<i>xdlcg</i>	0,171	0,186	0,17	0,167	0,176
<i>xclcg</i>	0,160	0,172	0,169	0,168	0,181
<i>xslcg</i>	0,23	0,245	0,236	0,225	0,25
método base					
<i>Sum</i> ⁺ <i>wTS</i>	0,364				

Com a amostra revisada, observa-se que o comportamento da estratégia Sum^+wTS^+GC no uso de cada tipo de grafo é o mesmo que o apresentado na amostra ruidosa. A estratégia utilizando os grafos abrangentes apresenta maior impacto no desempenho quando comparado ao impacto causado pelos grafos limitadores. Por exemplo, no subconjunto *menor*, as combinações *xdlcg+pr*, *xflcg+pr*, *dlcg+pr* apresentaram desempenhos de 0.35, 0.38, 0.40, respectivamente, e as combinações *slcg+out*, e *xslcg+out* apresentaram um desempenho de 0.50 e 0.51. Como os termos oriundos da coocorrência tendem a ganhar destaque no conjunto de recomendação, o Sum^+wTS^+GC potencializa a relevância destes termos. Esse comportamento é observado também nos demais subconjuntos de etiquetas,

ver Tabelas 4.7 e 4.8.

Tabela 4.7: Impacto no desempenho da estratégia Sum^+wTS^+GC pelo uso dos grafos abrangentes.

Subconjunto de etiquetas	Sum^+wTS^+GC	Desempenho
<i>menor</i>	$xdlcg+pr$, $xflcg+pr$, $dlcg+pr$	0.35, 0.38, 0.40
<i>misto</i>	$xdlcg+pr$, $xdlcg+out$, $dlcg+out$	0.23, 0.24, 0.26
<i>médio</i>	$xflcg+ind$, $xdlcg+ind$, $xdlcg+pr$	0.29, 0.27, 0.25
<i>maior</i>	$dlcg+out$, $clcg+out$, $xclcg+out$	0.15, 0.16, 0.16

Tabela 4.8: Impacto no desempenho da estratégia Sum^+wTS^+GC pelo uso dos grafos limitadores.

Subconjunto de etiquetas	Sum^+wTS^+GC	Desempenho
<i>menor</i>	$slcg+out$, e $xslcg+out$	0.50 e 0.51
<i>misto</i>	$slcg+out$, e $xslcg+out$	0.40 e 0.39
<i>médio</i>	$slcg+ind$, e $xslcg+ind$	0.41 e 0.41
<i>maior</i>	$slcg+out$, e $xslcg+auth$	0.24 e 0.25

Nas avaliações que seguem verificou-se o desempenho do Sum^+wTS^+GC na ausência de etiquetas nos vídeos que caracteriza o estado de *cold start*. Neste estado o conjunto final de recomendações não é impactado com os valores originários da coocorrência. Este fato torna a heurística de presença de termos nos atributos influente no ranqueamento dos candidatos a recomendação, a consequência é que os termos *spammers* se tornam ainda mais prejudiciais as recomendações, pois conta-se apenas com os termos dos atributos textuais do vídeo para realizar a recomendação.

Na amostra Youtube_Sports_3000, o Sum^+wTS^+GC utilizando alguns grafos de contexto são menos impactados pelo *cold start*, que o Sum^+wTS . Isto ocorre porque o processo da recomendação passa a integrar uma valoração dos termos também baseada pelo conhecimento das ligações entre os termos em uma base colaborativa, e não apenas pela presença de um termo nos atributos textuais, ver tabelas 4.5 e 4.11.

No subconjunto *menor*, a estratégia utilizando as combinações $xdlcg+ind$, e $xdlcg+pr$ apresentou uma queda de 18,5% no desempenho de suas recomendações, por outro lado o método base de comparação sofre um impacto de 28% em seu desempenho. Com uma quantidade mediana de etiquetas, o desempenho da estratégia utilizando as combinações $xclcg+ind$, $xclcg+out$, e $dlcg+out$ tem uma queda de 9,5%-12% e de 39% para a

método base. No subconjunto *maior*, a estratégia apresenta um resultado contrário aos demais subconjuntos citados, o desempenho chega a apresentar um aumento de 39,5%, na combinação *xslcg+out*, assim como o método base que obtêm um aumento de 20%, ver Tabela 4.11.

Nas Tabelas 4.9 e 4.10, observa-se que em todos os subconjuntos de etiquetas, os grafos abrangentes apresentam maior impacto no desempenho do Sum^+wTS^+GC do que os grafos limitadores, por valorarem mais termos no conjunto de recomendação. Por exemplo, no subconjunto *menor*, a estratégia usando o *xdlcg+pr*, *xclcg+pr*, *dlcg+pr* apresentou desempenhos de 0.23, 0.24, 0.24, respectivamente, já utilizando as combinações *slcg+out*, e *xslcg+out* apresentou desempenho de 0.29. Com a presença dos ruídos os grafos tendem a possuir mais ligações, porém com conceitos que não descrevem o conteúdo do vídeo, o que torna o ranqueamento do pagerank o que mais valora dentre os algoritmos de análise de *links*, e que conseqüentemente cause maior impacto do desempenho da estratégia quando está associado aos grafos abrangentes, como pode ser observado nos demais subconjuntos de etiquetas.

Tabela 4.9: Impacto no desempenho da estratégia Sum^+wTS^+GC pelo uso dos grafos abrangentes.

Subconjunto de etiquetas	Sum^+wTS^+GC	Desempenho
<i>menor</i>	<i>xdlcg+pr</i> , <i>xclcg+pr</i> , <i>dlcg+pr</i>	0.23, 0.24, 0.24
<i>misto</i>	<i>xflcg+pr</i> , <i>xdlcg+pr</i> , <i>dlcg+pr</i>	0.21, 0.22, 0.23
<i>médio</i>	<i>xflcg+pr</i> , <i>xdlcg+pr</i> , <i>dlcg+pr</i>	0.23, 0.24, 0.25
<i>maior</i>	<i>dlcg+pr</i> , <i>xflcg+pr</i> , <i>xdlcg+pr</i>	0.25, 0.25, 0.23

Tabela 4.10: Impacto no desempenho da estratégia Sum^+wTS^+GC pelo uso dos grafos limitadores.

Subconjunto de etiquetas	Sum^+wTS^+GC	Desempenho
<i>menor</i>	<i>slcg+out</i> , e <i>xslcg+out</i>	0.29
<i>misto</i>	<i>slcg+out</i> , e <i>xslcg+out</i>	0.34 e 0.35
<i>médio</i>	<i>slcg+out</i> , e <i>xslcg+out</i>	0.28 e 0.29
<i>maior</i>	<i>slcg+out</i> , e <i>xslcg+auth</i>	0.49 e 0.50

Tabela 4.11: Desempenho do método Sum^+wTS^+GC com a amostra Youtube_Sports_3000 em estado *cold start*

Subconjunto Menor de etiquetas					
Método proposto:	Algoritmos de análise de <i>links</i>				
Sum^+wTS^+GC	outdegree (out)	indegree (ind)	pagerank (pr)	hub (hub)	authority (auth)
<i>flcg</i>	0,268	0,271	0,246	0,261	0,271
<i>dlcg</i>	0,265	0,268	0,241	0,261	0,271
<i>clcg</i>	0,278	0,281	0,248	0,261	0,271
<i>slcg</i>	0,288	0,288	0,278	0,280	0,280
<i>xflcg</i>	0,272	0,278	0,243	0,274	0,2849
<i>xdlcg</i>	0,269	0,275	0,234	0,274	0,2849
<i>xclcg</i>	0,272	0,284	0,242	0,264	0,277
<i>xslcg</i>	0,293	0,293	0,283	0,285	0,2858
método base					
Sum^+wTS	0,293				
Subconjunto Misto de etiquetas					
Método proposto:	Algoritmos de análise de <i>links</i>				
Sum^+wTS^+GC	outdegree (out)	indegree (ind)	pagerank (pr)	hub (hub)	authority (auth)
<i>flcg</i>	0,270	0,266	0,235	0,268	0,273
<i>dlcg</i>	0,268	0,263	0,226	0,268	0,273
<i>clcg</i>	0,272	0,278	0,234	0,272	0,270
<i>slcg</i>	0,347	0,327	0,294	0,328	0,303
<i>xflcg</i>	0,265	0,265	0,220	0,273	0,284
<i>xdlcg</i>	0,259	0,269	0,217	0,273	0,283
<i>xclcg</i>	0,257	0,277	0,236	0,267	0,273
<i>xslcg</i>	0,352	0,332	0,298	0,334	0,309
método base					
Sum^+wTS	0,351				
Subconjunto Médio de etiquetas					
Método proposto:	Algoritmos de análise de <i>links</i>				
Sum^+wTS^+GC	outdegree (out)	indegree (ind)	pagerank (pr)	hub (hub)	authority (auth)
<i>flcg</i>	0,269	0,268	0,253	0,266	0,267
<i>dlcg</i>	0,268	0,268	0,246	0,266	0,267
<i>clcg</i>	0,271	0,270	0,257	0,267	0,267
<i>slcg</i>	0,283	0,281	0,274	0,280	0,278
<i>xflcg</i>	0,272	0,268	0,236	0,283	0,277
<i>xdlcg</i>	0,275	0,266	0,231	0,283	0,277
<i>xclcg</i>	0,273	0,271	0,258	0,273	0,271
<i>xslcg</i>	0,288	0,286	0,279	0,285	0,283
método base					
Sum^+wTS	0,285				
Subconjunto Maior de etiquetas					
Método proposto:	Algoritmos de análise de <i>links</i>				
Sum^+wTS^+GC					

	outdegree (out)	indegree (ind)	pagerank (pr)	hub (hub)	authority (auth)
<i>flcg</i>	0,340	0,338	0,284	0,331	0,326
<i>dlcg</i>	0,336	0,322	0,250	0,331	0,326
<i>clcg</i>	0,344	0,346	0,274	0,331	0,316
<i>slcg</i>	0,499	0,474	0,419	0,487	0,458
<i>xflcg</i>	0,330	0,318	0,248	0,330	0,330
<i>xdlcg</i>	0,315	0,318	0,235	0,330	0,330
<i>xclcg</i>	0,321	0,355	0,274	0,328	0,323
<i>xslcg</i>	0,505	0,482	0,423	0,495	0,466
método base					
<i>Sum⁺wTS</i>	0,521				

Na ausência de etiquetas, os resultados apresentados pelo *Sum⁺wTS⁺GC* na amostra Youtube_Sports_Rev em relação a amostra Youtube_Sports_3000 mostram que a remoção dos ruídos proporciona uma melhoria no desempenho nos subconjuntos de etiquetas *menor*, *médio* e *misto*, e que os maiores aumentos nos desempenhos da estratégia foram obtidos usando os grafos expandidos, ver tabelas 4.11 e 4.12.

Como o conjunto de recomendação não possui em sua composição os termos oriundos da coocorrência, os termos valorados pelos grafos, principalmente as etiquetas formais adicionados pelos grafos expandidos ganham relevância no ranqueamento, fazendo este tipo de grafo mais efetivo em *cold start*. No subconjunto *menor*, o *Sum⁺wTS⁺GC* utilizando o *clcg+ind*, *flcg+auth*, *dlcg+auth* e *clcg+auth* apresentou um aumento de cerca de 30%-32% no desempenho, o menor dentre os grafos avaliados, enquanto que o maior aumento obtido foi de cerca de 41% pelas combinações *xflcg+ind* e *xdlcg+ind*. No subconjunto *médio*, o *Sum⁺wTS⁺GC* utilizando os grafos não expandidos apresentou um aumento de cerca de 22%-23% no desempenho, enquanto que utilizando grafos expandidos *xdlcg+out*, *xclcg+out* e *xclcg+pr* obteve um aumento de 26%-29%. No subconjunto *misto*, o *xflcg+ind* e o *xdlcg+out* apresentam um aumento de 36% no desempenho, enquanto que o menor aumento obtido é de entre 16%-20% pelas combinações do grafo *slcg*.

No subconjunto *maior*, ocorreu um aumento de 11-18% na estratégia usando grafos abrangentes, como os *xflcg+ind*, *xdlcg+ind* e *xdlcg+pr*. E uma redução de 3%-13% utilizando grafos limitadores, como *slcg* e *xslcg*. Grafos abrangentes possuem um maior número de termos do que os grafos limitadores, com a remoção dos termos ruídos, o impacto exercido por esse tipo de grafo é maior no ranqueamento dos termos candidatos a recomendação. Além disso, a maior quantidade de termos candidatos a recomendação pontencializa a escolha dos grafos limitadores por apenas um subgrafo para valoração, característica mencionada anteriormente.

Tabela 4.12: Desempenho do método Sum^+wTS^+GC com a amostra Youtube_Sports_Rev em estado de *cold start*

Subconjunto Menor de etiquetas					
Método proposto:	Algoritmos de análise de <i>links</i>				
Sum^+wTS^+GC	outdegree (out)	indegree (ind)	pagerank (pr)	hub (hub)	authority (auth)
<i>flcg</i>	0,4075	0,4025	0,385	0,392	0,397
<i>dlcg</i>	0,4	0,4	0,377	0,392	0,397
<i>clcg</i>	0,402	0,4	0,382	0,3925	0,397
<i>slcg</i>	0,42	0,417	0,412	0,417	0,415
<i>xflcg</i>	0,462	0,47	0,417	0,455	0,462
<i>xdlcg</i>	0,457	0,47	0,397	0,455	0,462
<i>xclcg</i>	0,445	0,447	0,427	0,44	0,445
<i>xslcg</i>	0,467	0,465	0,46	0,465	0,462
método base					
Sum^+wTS	0,42				
Subconjunto Misto de etiquetas					
Método proposto:	Algoritmos de análise de <i>links</i>				
Sum^+wTS^+GC	outdegree (out)	indegree (ind)	pagerank (pr)	hub (hub)	authority (auth)
<i>flcg</i>	0,374	0,366	0,349	0,366	0,373
<i>dlcg</i>	0,372	0,365	0,332	0,366	0,373
<i>clcg</i>	0,375	0,368	0,344	0,372	0,371
<i>slcg</i>	0,415	0,410	0,388	0,407	0,396
<i>xflcg</i>	0,404	0,415	0,360	0,412	0,415
<i>xdlcg</i>	0,404	0,413	0,327	0,412	0,415
<i>xclcg</i>	0,395	0,398	0,376	0,397	0,401
<i>xslcg</i>	0,444	0,439	0,417	0,434	0,423
método base					
Sum^+wTS	0,42				
Subconjunto Médio de etiquetas					
Método proposto:	Algoritmos de análise de <i>links</i>				
Sum^+wTS^+GC	outdegree (out)	indegree (ind)	pagerank (pr)	hub (hub)	authority (auth)
<i>flcg</i>	0,348	0,347	0,335	0,335	0,342
<i>dlcg</i>	0,348	0,344	0,322	0,335	0,342
<i>clcg</i>	0,345	0,347	0,330	0,337	0,342
<i>slcg</i>	0,366	0,362	0,353	0,364	0,356
<i>xflcg</i>	0,389	0,378	0,336	0,381	0,382
<i>xdlcg</i>	0,385	0,372	0,305	0,381	0,382
<i>xclcg</i>	0,374	0,368	0,350	0,363	0,365
<i>xslcg</i>	0,393	0,389	0,379	0,390	0,383
método base					
Sum^+wTS	0,369				

Subconjunto Maior de etiquetas					
Método proposto:	Algoritmos de análise de <i>links</i>				
Sum^+wTS^+GC	outdegree (out)	indegree (ind)	pagerank (pr)	hub (hub)	authority (auth)
<i>flcg</i>	0,355	0,354	0,317	0,344	0,349
<i>dlcg</i>	0,35	0,35	0,269	0,344	0,349
<i>clcg</i>	0,353	0,354	0,291	0,348	0,346
<i>slcg</i>	0,433	0,429	0,403	0,43	0,42
<i>xflcg</i>	0,345	0,361	0,302	0,352	0,36
<i>xdlcg</i>	0,342	0,358	0,288	0,352	0,36
<i>xclcg</i>	0,340	0,359	0,308	0,34	0,349
<i>xslcg</i>	0,445	0,442	0,412	0,440	0,427
método base					
Sum^+wTS	0,451				

Assim como na amostra Youtube_Sports_3000, a amostra Youtube_Sports_Rev apresenta grafos que sofrem um pouco menos com o *cold start* em relação ao Sum^+wTS . A diferença de desempenho variou entre 5%-8%, o que mostra que o conhecimento colaborativo extraído pela estratégia é benéfico para amenizar o impacto causado pela ausência de etiquetas nos vídeos, ver Tabelas 4.6 e 4.12. No subconjunto *menor*, a estratégia utilizando as combinações *xclcg+ind*, *xclcg+out*, *xdlcg+hub* apresentou uma queda de 3% e de 18% no método base. O impacto de 3%-5% foi apresentado pela estratégia subconjunto *médio* utilizando as combinações *flcg+out*, *flcg+ind*, *dlcg+ind*, e o método base apresentou uma queda de 17%. Enquanto que no subconjunto *maior*, assim como no estado sem *cold start*, apresenta um aumento de cerca de 48% com a estratégia utilizando o *xslcg+out*, e *xslcg+hub*, diante dos 19% de aumento obtido pelo método base.

Em relação ao desempenho da estratégia com os grafos na amostra Youtube_Sports_Rev, nos subconjuntos *menor*, *misto* e *médio*, o Sum^+wTS^+GC que utilizam os grafos expandidos apresentaram ganhos de 5%-11% em relação ao método base, ver Tabela 4.12. Em *cold start* os grafos extendidos obtiveram melhor resultado porque a adição dos termos da tabela de redirecionamentos favorecem o valor recebido pelos termos com a configuração obtida por esses grafos pelos algoritmos de análise de *links*. Principalmente a forma como os algoritmos indegree e outdegree adicionam os termos da tabela de redirecionamentos da Wikipédia ao conjunto de recomendação, favorecendo ou não as etiquetas formais dependendo da quantidade de conceitos que estão disponíveis para descrição de um vídeo.

No subconjunto *menor*, as combinações *xflcg+ind*, *xdlcg+ind*, e *xslcg+ind* apresentaram desempenhos de 0.47, 0.47, 0.46, respectivamente. Com esta configuração, o

Sum^+wTS^+GC é indicado para recomendações formais, pois o uso do indegree faz com que as etiquetas formais sejam favorecidas no conjunto de recomendação, ainda mais nesse subconjunto que tende a conter um quantidade menor de termos em seus atributos textuais.

Os subconjuntos *médio* e *misto* por terem uma maior quantidade de termos presentes no conjunto de recomendação, torna o Sum^+wTS^+GC utilizando os grafos expandidos combinados com o outdegree mais indicado para recomendação informal. O outdegree valora os termos informais, o que tende a posicionar as etiquetas formais, adicionados pelos grafos, abaixo dos termos oriundos dos atributos textuais. No subconjunto *misto*, as combinações $xslcg+out$, $xslcg+out$, e $xslcg+hub$ apresentaram desempenhos de 0.44, 0.44, 0.43, respectivamente. No subconjunto *médio*, as combinações $xslcg+out$, $xdlclg+out$, e $xflcg+out$ apresentaram desempenhos de 0.39.

Na Tabela 4.12, observa-se que o comportamento do Sum^+wTS^+GC com a amostra revisada é caracterizado pela versão dos grafos utilizados, em vez do comportamento caracterizado pelos tipos de grafos apresentado pelos resultados da amostra Youtube_Sports_3000. Nos subconjuntos de dados em que houve ganhos, observa-se que a estratégia utiliza os grafos expandidos, independentes do tipo.

Na Tabela 4.13, observa-se que apesar da diferença de desempenho entre os grafos expandidos não ser grande, destaca-se que no subconjunto *menor* o Sum^+wTS^+GC que utiliza os grafos mais abrangentes ($xflcg$, $xdlclg$) tem uma tendência de terem maior desempenho do que quando utiliza o grafo limitador ($xslcg$), por exemplo o $xflcg+ind$, $xdlclg+ind$, e $xslcg+ind$ apresentaram desempenhos de 0.47, 0.47, 0.46, respectivamente.

Essa situação inverte a medida que o número de etiquetas cresce, pois nos subconjuntos de etiquetas *médio* e *misto* o Sum^+wTS^+GC utilizando o grafo $xslcg$ apresenta maior impacto que os grafos abrangentes. Como a métrica avalia o desempenho por meio das cinco primeiras etiquetas recomendadas, a medida que o número de etiquetas cresce o grafo $xslcg$ torna-se mais efetivo.

Tabela 4.13: Impacto no desempenho da estratégia Sum^+wTS^+GC pelo uso dos grafos expandidos.

Subconjunto de etiquetas	Sum^+wTS^+GC	Desempenho
<i>menor</i>	$xflcg+ind$, $xdlclg+ind$, e $xslcg+ind$	0.47, 0.47, 0.46
<i>misto</i>	$xslcg+out$, $xslcg+out$, e $xslcg+hub$	0.44, 0.44, 0.43
<i>médio</i>	$xslcg+out$, $xdlclg+out$, e $xflcg+out$	0.39

O comportamento do Sum^+wTS^+GC no subconjunto de etiquetas *maior* é diferente, como esse subconjunto tende a possuir um maior nível de descrição do conteúdo do vídeo, contendo termos mais formais nos atributos textuais, a diferença de desempenho da estratégia entre as versões dos grafos é menor. Por exemplo, as combinações *dlcg+pr*, *clcg+pr*, *xdlcg+pr* apresentaram desempenhos de 0.27, 0.29, 0.29, respectivamente. Esse fato aliado a quantidade de etiquetas consideradas boas pela métrica torna o desempenho da estratégia mais impactado entre os demais subconjuntos de etiquetas, pois não existe um equilíbrio no conjunto de recomendação entre as etiquetas oriundas dos atributos textuais e as etiquetas formais.

A estratégia obteve os menores impactos nas recomendações utilizando o *slcg*, pois não adiciona etiquetas formais novas e influencia o conjunto de recomendação com a valoração de um termo, principalmente quando está combinado com os algoritmos de análise de *links* indegree e outdegree, ver Tabela 4.14.

Tabela 4.14: Impacto no desempenho da estratégia Sum^+wTS^+GC pelo uso dos grafos limitadores.

Subconjunto de etiquetas	Sum^+wTS^+GC	Desempenho
<i>menor</i>	<i>slcg+out</i> , e o <i>slcg+ind</i>	0.42
<i>misto</i>	<i>slcg+out</i> , e o <i>xslcg+ind</i>	0.41
<i>médio</i>	<i>slcg+out</i> , e o <i>slcg+ind</i>	0.36
<i>maior</i>	<i>slcg+out</i> , e o <i>xslcg+out</i>	0.43, 0.44

4.4 Discussões Sobre o Capítulo 4

Neste capítulo caracterizou-se o conceito de grafos de contexto e sua função de busca dos termos significativos de um texto curto. Os atributos textuais mostram-se uma fonte alternativa na produção de termos candidatos a recomendação, porém o desempenho do método baseia-se na forma como os usuários fornecem as informações dos vídeos. Por isso é preciso estudar maneiras de minimizar o impacto causado pelos ruídos nas sugestões quando não existem informações prévias sobre o vídeo. Com a finalidade de atuar positivamente neste fator na recomendação que apresentou-se a estratégia de recomendação associativo com os grafos de contexto, partindo para uma análise da maneira como cada grafo pode contribuir para o ranqueamento dos termos de acordo com determinado tipo de recomendação.

Assim como o método Sum^+wTS apresentou um grande impacto em seu desempe-

nho devido as características da amostra Youtube_Sports mencionadas no Capítulo 3, na primeira rodada dos experimentos a estratégia Sum^+wTS^+GC também apresentou baixo desempenho devido essas características. Os desempenhos não mostram o real desempenho dos grafos e por esta razão realizou-se uma nova rodada de experimentos com essa amostra sendo revisada.

Na segunda rodada de experimentos verifica-se primeiro o comportamento dos grafos no Sum^+wTS^+GC na presença de etiquetas dos vídeos que compõem as amostras Youtube_Sports_3000 e Youtube_Sports_Rev. Destaca-se que o comportamento dos grafos abrangentes e limitadores foi identificado neste estado de ausência de etiquetas. O Sum^+wTS^+GC usando os grafos abrangentes é ideal para recomendações mais conceituais, principalmente quando utilizam os algoritmos indegree e hub, enquanto que utilizando os grafos limitadores é ideal para recomendações que não demandam um certo nível de formalismo.

Em geral o Sum^+wTS^+GC usando os grafos simples apresentou uma função de refinamento do conjunto de recomendações por considerar a presença do termo como artigo e ligações na Wikipédia relevante. A expansão dos grafos simples apresentou-se com um recurso que fornece a estratégia, uma recomendação com etiquetas formais originárias dos termos extraídos dos atributos textuais. Entretanto esse recurso disputa com o peso dos valores oriundos da coocorrência, o que torna os resultados obtidos na amostra menor que o método base.

Em seguida, avaliou-se o desempenho da estratégia proposta nas amostras Youtube_Sports_3000 e Youtube_Sports_Rev na ausência de etiquetas. Verificou-se que este recurso mostrou-se importante em *cold start* pois cria-se a possibilidade de tornar o conjunto de recomendação mais formal em relação ao significado, o que favorece a formação de uma coleção mais consistente. O Sum^+wTS^+GC utilizando a maioria dos grafos expandidos apresentou ganho em desempenho em relação ao método base, principalmente em vídeos que apresentam de três a nove etiquetas. Com a ausência da coocorrência, as etiquetas formais ganham mais relevância nos conjuntos de recomendação, sendo mais efetivas no subconjunto de etiquetas menor, principalmente utilizando o algoritmo indegree. Entretanto, essa eficiência diminui à medida que os vídeos apresentam mais etiquetas, sendo a estratégia utilizando a combinação do grafo limitador mais o algoritmo de análise de *links* outdegree mais eficiente. Porém estes desempenhos estão diretamente relacionados com as cinco primeiras etiquetas consideradas relevantes pela métrica $P@n$, sendo natural o fato do desempenho do Sum^+wTS^+GC cair com o aumento da quantidade de etiquetas em cada subconjunto de etiquetas.

Capítulo 5

Avaliação qualitativa da estratégia de recomendação associativa de etiquetas usando grafos de contexto

Neste capítulo mostram-se o ambiente de experimentação e os resultados de uma análise qualitativa das etiquetas recomendadas pela estratégia Sum^+wTS^+GC .

A integração dos grafos de contexto no Sum^+wTS tem como objetivo a identificação de conceitos mais amplos que aqueles expressos nos atributos textuais dos vídeos, permitindo o controle do nível de formalismo no conjunto de termos candidatos a recomendação. Essa propriedade contribui para a redução do impacto da ambiguidade e dos erros de grafia, que são recorrentes em recomendação de termos, principalmente em aplicações colaborativas.

Na avaliação automática, discutida no Capítulo 4, apresenta-se o desempenho do Sum^+wTS^+GC considerando o uso de cada grafo de contexto e sua associação com os algoritmos de análise de *links*, sendo o nível de formalismo, resultante dessa associação, um aspecto importante para a análise do desempenho da estratégia.

Observa-se que o nível de formalismo influencia positivamente no desempenho da estratégia em vídeos com uma quantidade pequena e mediana de etiquetas, mas o comportamento das pessoas diante de diferentes modos de recomendação, permitidos pela estratégia, não pode ser analisada pela métrica de avaliação utilizada. A avaliação qualitativa, apresentada neste capítulo, complementa o estudo realizado usando a avaliação automática, através da análise das escolhas realizadas pelos avaliadores. A decisão de aceitar os termos mais conceituais nas recomendações fica a cargo do nível de formalismo que uma aplicação possa demandar. Neste capítulo opta-se, pela análise do comportamento

dos avaliadores ao tipo de recomendação demandado pela aplicação.

Os experimentos realizados capturam o comportamento dos avaliadores diante de dois tipos de recomendações: informal e formal. A captura do comportamento é realizada pela análise dos seguintes fatores: o perfil dos avaliadores, a cobertura dos termos na Wikipédia, a presença de termos formais nas recomendações e as posições que estas etiquetas ocupam nas sugestões.

Este capítulo organiza-se da seguinte forma: na Seção 5.1 apresenta-se o ambiente desenvolvido para os experimentos. Na Seção 5.2 apresenta-se a análise dos resultados obtidos nos experimentos. Por fim, na Seção 5.3 estão as considerações finais deste capítulo.

5.1 Ambiente de experimentação

Na Figura 5.1, apresenta-se o ambiente de experimentação construído para a condução dos estudos qualitativos.

Os módulos de pré-processamento e de métodos são os mesmos usados na avaliação automática. O módulo avaliação usa uma aplicação web para capturar a percepção de pessoas quanto as recomendações das estratégias. Os *scripts* desse módulo, além de obterem o desempenho de cada estratégia, também extraem informações sobre as avaliações das recomendações.

Para participar das avaliações, os voluntários precisaram realizar o cadastro na aplicação web, de onde se construiu um perfil mínimo, com informações sobre gênero, idade, formação acadêmica e a relação que possui com o tema esportivo, que é autodeclarado: expert, acompanha as finais, não têm interesse no tema.

Cada vez que o avaliador acessa a aplicação, o sistema exibe cinco vídeos, escolhidos de forma aleatória, de até três minutos de duração. Ao final de cada vídeo, a aplicação exibe o conjunto de etiquetas recomendadas em um campo. Em seguida, o avaliador deve escolher as etiquetas que mais descrevem o vídeo reproduzido.

As escolhas feitas são armazenadas no banco de dados e posteriormente analisadas segundo a métrica $P@n$, e a extração de informações das etiquetas escolhidas são coletadas.

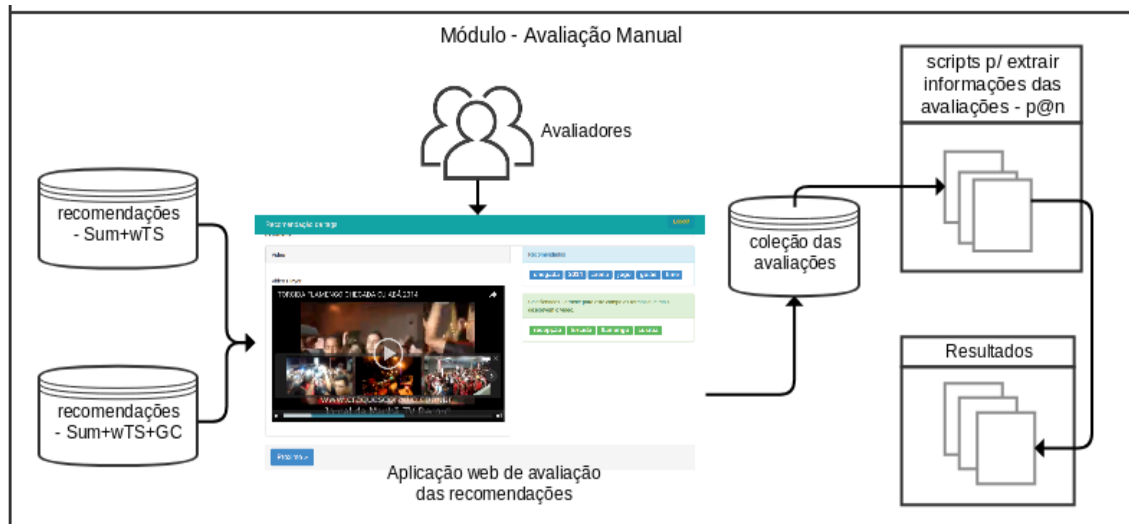


Figura 5.1: Ambiente de experimentação.

Os avaliadores que participaram do experimento era em sua maioria jovens, com ensino superior. Em relação a afinidade com o tema esportivo, 31% dos avaliadores não acompanham esportes, 37% assistem somente as finais, e 32% consideram-se experts no tema esporte, ver tabela 5.1.

Tabela 5.1: Perfil dos avaliadores

Afinidade com o tema esporte		
Não acompanha	finais	experts
35	42	36
Escolaridade		
Ensino Médio	Superior Incompleto	Superior completo
1	58	68
Faixa etária		
18-25 anos	26-35 anos	36-45 anos
73	32	7

5.2 Resultados numéricos

Nesta seção apresentam-se os resultados obtidos para uma avaliação qualitativa das recomendações realizadas pelos métodos Sum^+wTS e $Sum^+wTS+GC$ na ausência de etiquetas nos vídeos. Foram selecionados apenas quatro grafos de contexto, aqueles que apresentaram desempenho superior ao método definido como base para esta pesquisa

(Sum^+wTS). O Sum^+wTS^+GC composto do grafo de contexto e algoritmo de análise de *links*, $xflcg+out$, e $xdlcg+out$ foi avaliado no subconjunto de etiquetas *menor*, e as composições $xflcg+ind$ e $xslcg+ind$, foi avaliado usando o subconjunto de etiquetas *médio*.

Diferente dos subconjuntos de etiquetas apresentados no Capítulo 3, que são compostos pelos metadados dos vídeos, os subconjuntos de etiquetas menor e médio usados neste estudo, são formados pelas recomendações geradas de 400 vídeos únicos que foram selecionados da amostra Youtube_Sports_Rev. Para o método base a amostra é composta por 400 recomendações, sendo 200 para cada subconjunto (*menor* e *médio*). Enquanto que para o método Sum^+wTS^+GC a amostra é composta por 800 recomendações, sendo 200 recomendações para cada grafo em seus respectivos subconjuntos.

A avaliação da qualidade das recomendações envolvendo pessoas revela o quanto elas estão dispostas a aceitar o tipo informal ou formal de recomendação que uma aplicação possa oferecer. Nos dois subconjuntos analisados, o Sum^+wTS^+GC adicionou etiquetas formais no conjunto de recomendação, com o uso dos grafos expandidos, e termos que apresentam maior relação entre si por meio da análise das ligações apresentados por eles na Wikipédia.

Na Tabela 5.2, mostra-se o desempenho do Sum^+wTS^+GC , obtido através do uso da métrica $P@n$ sobre as escolhas dos avaliadores. Observa-se que a estratégia apresenta um impacto no desempenho 8% menor no subconjunto de etiquetas *menor* (utilizando a combinação $xflcg+out$) e 10% menor no subconjunto de dados *médio* (utilizando a combinação $xslcg+ind$), quando comparado com o desempenho do método base Sum^+wTS .

No subconjunto de etiquetas *menor*, o Sum^+wTS^+GC utilizando os grafos $xflcg$ e $xdlcg$ valoram seus termos a partir do algoritmo outdegree, que favorece os termos informais no conjunto dos termos candidatos a recomendação. Desta maneira, o uso dessa composição é indicada para recomendações que não demandam representatividade conceitual nas etiquetas.

No subconjunto de etiquetas *médio*, o Sum^+wTS^+GC com os grafos de contexto $xflcg$ e $xslcg$ valoram seus termos a partir do algoritmo indegree, que favorece os termos mais formais no conjunto de recomendações. A estratégia utilizando esses grafos é indicada para recomendações que preocupam-se com a representatividade conceitual dos termos.

Tabela 5.2: Desempenho na avaliação qualitativa - Subconjuntos de etiquetas *menor* e *médio*

Subconjunto de etiquetas <i>menor</i>		
Método		desempenho
Sum^+wTS^+GC	$xflcg+out$	0,323
	$xdlcg+out$	0,350
Sum^+wTS		0,42
Subconjunto de etiquetas <i>médio</i>		
Método		desempenho
Sum^+wTS^+GC	$xflcg+ind$	0,219
	$xslcg+ind$	0,20
Sum^+wTS		0,282

Os resultados apresentados a seguir analisam os fatores que contribuíram para as escolhas dos avaliadores em relação as etiquetas recomendadas, principalmente em relação a adição de etiquetas formais ao conteúdo dos vídeos assistidos e como isso refletem nos resultados apresentados pelos grafos de contexto analisados. Os fatores analisados incluem a quantidade de etiquetas escolhidas, o perfil dos avaliadores, as posições das etiquetas formais nas recomendações, a composição dos conjuntos de recomendação e das escolhas.

5.2.1 Quais fatores e como eles influenciam a recomendação

Em geral os avaliadores escolheram em média 50% das etiquetas recomendadas pelos grafos de contexto em ambos os subconjuntos de etiquetas. Desse total, cerca de 72% estão entre as cinco primeiras no conjunto de recomendação. No método Sum^+wTS , 65% das etiquetas sugeridas são escolhidas pelos avaliadores e 80% delas estão presentes nas cinco primeiras posições no conjunto de recomendação, ver Tabela 5.3.

Tabela 5.3: Taxas de etiquetas escolhidas (taxa 1) e de etiquetas escolhidas que estão entre as cinco primeiras recomendadas (taxa 2).

Subconjunto de etiquetas menor			
Estratégia		taxa 1	taxa 2
Sum^+wTS^+GC	$xflcg+out$	54%	73%
	$xdlcg+out$	51%	74%
Sum^+wTS		67%	82%
Subconjunto de etiquetas médio			
Estratégia		taxa 1	taxa 2
Sum^+wTS^+GC	$xflcg+ind$	54%	61%
	$xslcg+ind$	50%	72%
Sum^+wTS		64%	80%

A análise desses resultados sugere que a posição que o termo formal ocupa, na recomendação influencia a sua escolha. Antes de realizarmos essa análise, verificou-se o desempenho da estratégia de acordo com cada grafo e perfil de avaliador.

Perfil dos avaliadores

Em geral, o desempenho do Sum^+wTS^+GC nas avaliações sobre o ponto de vista do tipo do avaliador não apresentou uma grande diferença entre os resultados. Os experts apresentaram desempenho 4% maior que os avaliadores que assistem apenas as finais e os que não acompanham esportes, ver Tabela 5.4.

Tabela 5.4: Desempenho do Sum^+wTS^+GC nas avaliações de acordo com o perfil do avaliador:

experts	finais	não acompanha
0.280	0.269	0.269

Os resultados apresentados na Tabela 5.5 foram obtidos por meio do gabarito extraído da amostra Youtube_Sports_Rev, utilizada no experimento descrito na Seção 4.3.2. A diferença nos resultados apresentados revela a relação entre o desempenho apresentado pela estratégia e a disponibilidade de etiquetas formais para os avaliadores. Observa-se, que mesmo com a diferença entre os cenários, em todos os perfis dos avaliadores, o Sum^+wTS^+GC utilizando o indegree apresenta uma queda de desempenho, o que reforça a existência da relação da presença das etiquetas formais e o desempenho.

Tabela 5.5: Desempenho do Sum^+wTS^+GC nas avaliações de acordo com o perfil do avaliador (em cada grafo) :

subconjuntos de etiquetas	Sum^+wTS^+GC	perfis		
		experts	finais	não acompanha
<i>menor</i>	<i>xflcg+out</i>	0.381	0.297	0.310
	<i>xdlcg+out</i>	0.377	0.329	0.358
<i>médio</i>	<i>xflcg+ind</i>	0.208	0.237	0.211
	<i>xslcg+ind</i>	0.205	0.203	0.192

Termos & artigos Wikipédia

Em relação a parcela de etiquetas que possuem artigos Wikipédia e que também compõem o conjunto de recomendação, o vocabulário apresentado pelos vídeos esportivos estão presentes no contexto da Wikipédia, apresentando uma taxa superior a 80% nos subconjuntos *menor* e *médio*, ver Tabela 5.6. A mesma situação é encontrada no conjunto de etiquetas escolhidas. Cerca de 80% das etiquetas escolhidas possuem artigos Wikipédia, ver Tabela 5.7. Isso indica que uma parcela significativa dos termos candidatos a recomendação tendem a apresentar uma influência na valoração obtida pelas ligações existentes entre esses termos (artigos) na Wikipédia, favorecendo a busca pelos termos candidatos que mais apresentam significado.

Tabela 5.6: Taxa de etiquetas sugeridas que possuem artigo na Wikipédia:

subconjuntos de etiquetas	Sum^+wTS^+GC	perfis		
		experts	finais	não acompanha
<i>menor</i>	<i>xflcg+out</i>	84%	88%	90%
	<i>xdlcg+out</i>	88%	85%	85%
<i>médio</i>	<i>xflcg+ind</i>	91%	91%	92%
	<i>xslcg+ind</i>	87%	86%	91%

Tabela 5.7: Taxa de etiquetas selecionadas que possuem artigo Wikipédia:

subconjuntos de etiquetas	$Sum^{+}wTS^{+}GC$	perfis		
		experts	finais	não acompanha
<i>menor</i>	$xflcg+out$	85%	87%	89%
	$xdlcg+out$	91%	85%	80%
<i>médio</i>	$xflcg+ind$	92%	89%	94%
	$xslcg+ind$	89%	87%	93%

Etiquetas formais

O $Sum^{+}wTS^{+}GC$ com o algoritmo de análise de *links* indegree (ind) é a composição indicada para produzir um conjunto de recomendação que contenha etiquetas formais. Na Tabela 5.8 pode-se ver que essa composição possui cerca de 30%-50% mais recomendações com etiquetas formais do que o $Sum^{+}wTS^{+}GC$ utilizando a composição dos grafos com o algoritmo de análise de *links* outdegree (out).

Tabela 5.8: Taxa de recomendações que possuem etiquetas formais:

subconjuntos de etiquetas	$Sum^{+}wTS^{+}GC$	perfis		
		experts	finais	não acompanha
<i>menor</i>	$xflcg+out$	26/55 - 47%	45/106 - 42%	14/29 - 48%
	$xdlcg+out$	16/53 - 30%	48/91 - 52%	21/46 - 45%
<i>médio</i>	$xflcg+ind$	69/83 - 83%	62/71 - 87%	31/39 - 79%
	$xslcg+ind$	42/63 - 66%	62/91 - 68%	30/39 - 76%

Em relação a composição dos conjuntos de recomendações, observa-se que de dez etiquetas sugeridas, três são etiquetas formais nas recomendações produzidas pelo $Sum^{+}wTS^{+}GC$ utilizando a composição $xflcg_ind$. O $Sum^{+}wTS^{+}GC$ utilizando os grafos com o algoritmo de análise de *links* outdegree possuem cerca de 20% etiquetas formais compondo as sugestões, ver Tabela 5.9.

Tabela 5.9: Taxa de etiquetas formais que foram sugeridas (compõem o conjunto de recomendação):

subconjuntos de etiquetas	Sum^+wTS^+GC	perfis		
		experts	finais	não acompanha
<i>menor</i>	<i>xflcg+out</i>	20%	24%	23%
	<i>xdlcg+out</i>	21%	23%	21%
<i>médio</i>	<i>xflcg+ind</i>	30%	29%	33%
	<i>xslcg+ind</i>	14%	13%	14%

Entre as recomendações do Sum^+wTS^+GC que possuem etiquetas formais geradas pela composição dos grafos com o algoritmo de análise de *links indegree*, em cerca de 40% das avaliações encontram-se etiquetas formais dentre as escolhidas. Destaca-se os avaliadores do perfil que não tem afinidade com o tema esportivo, que em metade de suas avaliações possuem etiquetas formais. Enquanto que o Sum^+wTS^+GC usando a composição dos grafos com o algoritmo de análise de *links outdegree*, as etiquetas formais estão presentes em média 25% das avaliações, ver Tabela 5.10. O conjunto de etiquetas escolhidas pelos avaliadores é composto de cerca de 20%-30% de etiquetas formais, ver Tabela 5.11.

Tabela 5.10: Taxa de avaliações que tinham etiquetas formais:

subconjuntos de etiquetas	Sum^+wTS^+GC	perfis		
		experts	finais	não acompanha
<i>menor</i>	<i>xflcg+out</i>	6/26 - 23%	13/45 - 28%	4/14 - 28%
	<i>xdlcg+out</i>	1/16 - 6%	8/48 - 16%	5/21 - 23%
<i>médio</i>	<i>xflcg+ind</i>	32/69 - 46%	26/62 - 41%	16/31 - 51%
	<i>xslcg+ind</i>	10/42 - 23%	12/62 - 19%	5/30 - 16%

Tabela 5.11: Taxa de etiquetas formais que foram escolhidas (compõem o conjunto de etiquetas escolhidas):

subconjuntos de etiquetas	Sum^+wTS^+GC	perfis		
		experts	finais	não acompanha
<i>menor</i>	<i>xflcg+out</i>	26%	27%	34%
	<i>xdlcg+out</i>	25%	29%	27%
<i>médio</i>	<i>xflcg+ind</i>	27%	26%	28%
	<i>xslcg+ind</i>	22%	27%	20%

Posição das etiquetas formais

Os resultados nas tabelas 5.12, 5.13 e 5.14 mostram a taxa de etiquetas formais presentes nos cinco, seis e sete primeiros termos sugeridos na recomendação, respectivamente. Esses resultados estão relacionados com a escolha realizada dessas etiquetas mostrada nas Tabelas 5.15, 5.16 e 5.17. Percebe-se que os avaliadores tendem a escolher as etiquetas formais por terem mais representatividade conceitual do conteúdo do vídeo e por estarem localizadas nas primeiras posições da sugestão.

Em uma aplicação que pretende obter uma recomendação com etiquetas mais conceituais, o Sum^+wTS^+GC utilizando a composição dos grafos com o algoritmo de análise de *links* indegree são os mais indicados, pois proporcionam maior qualidade das etiquetas, devido a relevância dada as etiquetas que estão mais relacionados entre si e a presença de etiquetas formais, posicionando-os nas primeiras posições da recomendação. Essa composição da estratégia apresenta taxas de escolhas de 80% a 100% das etiquetas formais recomendadas independente do perfil apresentado pelos avaliadores.

Tabela 5.12: Taxa de etiquetas formais sugeridas que estão entre as cinco primeiras recomendadas (conjunto de recomendação):

subconjuntos de etiquetas	Sum^+wTS^+GC	perfis		
		experts	finais	não acompanha
<i>menor</i>	<i>xflcg+out</i>	19%	21%	21%
	<i>xdlcg+out</i>	16%	25%	14%
<i>médio</i>	<i>xflcg+ind</i>	72%	79%	74%
	<i>xslcg+ind</i>	95%	95%	96%

Tabela 5.13: Taxa de etiquetas formais sugeridas que estão entre as seis primeiras recomendadas (conjunto de recomendação):

subconjuntos de etiquetas	Sum^+wTS^+GC	perfis		
		experts	finais	não acompanha
<i>menor</i>	<i>xflcg+out</i>	36%	36%	45%
	<i>xdlcg+out</i>	34%	39%	37%
<i>médio</i>	<i>xflcg+ind</i>	77%	85%	78%
	<i>xslcg+ind</i>	96%	96%	97%

Tabela 5.14: Taxa de etiquetas formais sugeridas que estão entre as sete primeiras recomendadas (conjunto de recomendação):

subconjuntos de etiquetas	Sum^+wTS^+GC	perfis		
		experts	finais	não acompanha
<i>menor</i>	<i>xflcg+out</i>	50%	54%	57%
	<i>xdlcg+out</i>	51%	51%	59%
<i>médio</i>	<i>xflcg+ind</i>	81%	88%	82%
	<i>xslcg+ind</i>	98%	96%	97%

Quando opta-se pela produção de recomendações que não enfatizam etiquetas formais i.e o Sum^+wTS^+GC usando a combinação dos grafos com o algoritmo de análise de *links* outdegree, observa-se que nas avaliações em que estão presentes, cerca de 50% das etiquetas formais escolhidas não estão posicionados entre os cinco primeiros sugeridos, como no caso da composição *xdlcg_out*, ver Tabela 5.15. Esta taxa chega a alcançar 25% a 19% quando considera-se uma maior quantidade de etiquetas relevantes na sugestão, ver Tabelas 5.16, 5.17.

Tabela 5.15: Taxa de etiquetas formais escolhidas que estão entre as cinco primeiras sugeridas:

subconjuntos de etiquetas	Sum^+wTS^+GC	perfis		
		experts	finais	não acompanha
<i>menor</i>	<i>xflcg+out</i>	0%	15%	25%
	<i>xdlcg+out</i>	0%	50%	0%
<i>médio</i>	<i>xflcg+ind</i>	80%	98%	81%
	<i>xslcg+ind</i>	100%	100%	100%

Tabela 5.16: Taxa de etiquetas formais escolhidas que estão entre as seis primeiras sugeridas:

subconjuntos de etiquetas	Sum^+wTS^+GC	perfis		
		experts	finais	não acompanha
<i>menor</i>	<i>xflcg+out</i>	50%	23%	25%
	<i>xdlcg+out</i>	0%	75%	20%
<i>médio</i>	<i>xflcg+ind</i>	87%	98%	84%
	<i>xslcg+ind</i>	100%	100%	100%

Tabela 5.17: Taxa de etiquetas formais escolhidas que estão entre as sete primeiras sugeridas:

subconjuntos de etiquetas	Sum^+wTS^+GC	perfis		
		experts	finais	não acompanha
<i>menor</i>	<i>xflcg+out</i>	66%	50%	50%
	<i>xdlcg+out</i>	0%	81%	60%
<i>médio</i>	<i>xflcg+ind</i>	92%	100%	87%
	<i>xslcg_ind</i>	100%	100%	100%

5.3 Discussões Sobre o Capítulo 5

Neste capítulo caracterizou-se o ambiente de experimentação da avaliação qualitativa dos métodos Sum^+wTS e Sum^+wTS^+GC , e os resultados obtidos nos experimentos.

O estudo realizado na avaliação automática apresentou resultados que permitiram uma medição quantitativa do desempenho do Sum^+wTS^+GC perante as versões obtidas da combinação entre os tipos de grafos de contexto e os algoritmos de análise de *links*, entretanto esse tipo de avaliação não mostra como as pessoas reagem a determinado tipo de recomendação que a estratégia proporciona, com isso optou-se pela realização de uma avaliação qualitativa da estratégia. Dessa forma a avaliação qualitativa complementa o estudo realizado com a avaliação quantitativa do desempenho da estratégia, através da observação do comportamento dos avaliadores diante dos tipos de recomendações abordados.

As informações extraídas das avaliações confirmam a relação entre a presença de etiquetas formais e o desempenho apresentado pelos grafos. Quando a aplicação realiza recomendações que demandam um nível de formalismo em seus termos por meio do Sum^+wTS^+GC usando a composição dos grafos com o algoritmo de análise de *links* indegree, os avaliadores tenderam a escolher essas etiquetas independente do perfil, principalmente por eles estarem mais relacionados ao conteúdo do vídeo e estarem entre os primeiras recomendadas.

Observou-se que cerca de 70% das etiquetas escolhidas pelos avaliadores estão presentes entre as cinco primeiras recomendadas. Nas recomendações produzidas pelo Sum^+wTS^+GC utilizando a combinação dos grafos com o algoritmo de análise de *links* outdegree, as etiquetas formais tendem a ficar fora das cinco primeiras recomendadas. Entretanto, das avaliações que possuíam essas etiquetas, os avaliadores tenderam a escolhê-las pela sua re-

presentatividade conceitual, mesmo que a aplicação não demande um nível de formalismo, considerando um número maior de etiquetas relevantes no conjunto de recomendação.

A Wikipédia é um enciclopédia colaborativa. A versão em português, comparada a versão em inglês, possui um número menor de artigos e principalmente de páginas de redirecionamento que são responsáveis por proporcionar termos mais formais nas recomendações. Este fato reforça a importância de se ter uma coleção de dados bem abrangente ao conteúdo que se deseja associar as etiquetas, e que isto faz diferença no desempenho da estratégia.

Capítulo 6

Considerações Finais

As estratégias de recomendação de etiquetas que são baseadas em padrões de coocorrência com etiquetas previamente associadas ao objeto vêm consistentemente produzindo resultados estado da arte. O Sum^+wTS é um exemplo desse tipo de método de recomendação que é composto da coocorrência de termos em uma base de dados integrado a uma métrica de relevância baseada na presença dos termos nos atributos textuais do objeto. Entretanto, a eficácia desse método apresenta redução no estado onde o objeto não apresenta etiquetas previamente associadas, o que caracteriza o problema de *cold start*. Desta forma, um dos maiores desafios dos métodos associativos de etiquetagem, em especial o Sum^+wTS é recomendar termos relevantes quando esta nesse estado e têm-se um nível de informação limitada sobre o conteúdo do objeto alvo da recomendação.

Nesta dissertação buscou-se compreender o comportamento do método associativo de etiquetas Sum^+wTS na ausência de etiquetas previamente associadas aos vídeos, utilizando vídeos esportivos. Além disso, propôs-se uma estratégia que integre o método Sum^+wTS ao conceito de grafos de contexto, cujo objetivo é melhorar a recomendação de etiquetas no estado de *cold start*. A integração ocorre no sentido de permitir uma recomendação baseada não somente pela presença das etiquetas nos atributos textuais, mais associada ao conhecimento extraído das ligações dos conceitos presentes na Wikipédia e a possibilidade de uma aplicação escolher determinado nível de formalismo em seu processo de etiquetagem.

O estudo inicial foi dividido em duas partes. Realizou-se, inicialmente, uma validação da implementação do método base e um estudo do seu desempenho nos estados de ausência e presença de etiquetas nos vídeos, utilizando a amostra esportiva formada neste trabalho. Foi observada na amostra a presença de ruídos, como termos que não descreviam o real

conteúdo do vídeo, resultante da prática dos usuários na busca de dar relevância para o seus vídeos, e também a diversidades gramaticais presentes na língua portuguesa, que a torna mais complexa que a língua inglesa. O que impactou negativamente a medição do desempenho do método, sendo necessário a realização de uma revisão da amostra e uma nova rodada de experimentos. Após a revisão da amostra, os resultados apresentam o impacto esperado do problema de *cold start* nos subconjunto de dados definidos.

Na segunda parte, analisou-se o desempenho do Sum^+wTS^+GC dispondo de cada grafo de contexto apresentado, observando o comportamento do conjunto de recomendação de acordo com os tipos de grafos, suas versões e algoritmos de análise de links. Essa análise ocorreu no estado de ausência e presença de *cold start*. Os primeiros resultados apontam que os ruídos afetam consideravelmente o desempenho da estratégia em todos os subconjuntos de dados. Além disso, com a amostra revisada, observou-se que a estratégia apresentou melhoria no desempenho em vídeos com até uma quantidade mediana e mista de etiquetas, em relação ao método base. Também destaca-se os perfis de recomendação que podem ser utilizadas de acordo com a combinação dos tipos de grafos e os algoritmos de análise de *links*, permitindo a uma aplicação utilizar a estratégia de acordo com sua necessidade.

O segundo estudo foi realizado através de uma análise qualitativa das recomendações produzidas pela estratégia proposta em *cold start*. Como o estudo inicial focou na observação da influência dos grafos nas recomendações, e consequentemente identificação do uso dos grafos em um determinado tipo de recomendação, essa análise visa observar o comportamento dos avaliadores diante de perfis diferentes de recomendação em relação ao formalismo, com o intuito de complementar a análise realizada na primeira parte do estudo. Foi observado que as pessoas tendem a etiquetar um vídeo utilizando a união dos termos mais conectados entre si, oriundos do conhecimento obtido das ligações na Wikipédia e a adição das etiquetas formais, quando a recomendação proporcionada pela estratégia utilizando grafo que estimula a relevância dessas etiquetas no conjunto de recomendação é utilizada, sendo observada em todos os perfis de usuários definidos. Além disso, houve avaliadores que também escolheram etiquetas formais, mesmo quando não estão presentes entre as primeiras etiquetas recomendadas, caracterizando uma aplicação que não demanda um nível de formalismo em suas sugestões.

As maiores dificuldades encontradas durante este trabalho foram em relação ao processo de formação da coleção utilizada e a identificação dos ruídos presente na amostra. A construção da coleção demandou um período de tempo considerável, devido a especificidade da temática dos vídeos e a escolha de vídeos que tenham em sua maioria termos da

língua portuguesa, foram necessários inúmeras coletas, além do processamento dos dados obtidos ter sido um processo longo e seletivo. Enquanto que a presença de termos spamms obrigou a realização de uma revisão da amostra, que consistiu na verificação das etiquetas fornecidas estarem de acordo com o conteúdo apresentado pelos vídeos, o que demandou grande esforço.

Como trabalhos futuros pretende-se ainda fazer o aprimoramento do $Sum^{+}wTS^{+}GC$ para que possa ter a capacidade de simular o *feedback* do usuário. Além disso, a estratégia proposta estimula mais pesquisa no sentido de incorporar recursos que possam explorar fontes de termos a recomendação, visando a formação de grafos mais robustos, permitindo maiores participação na valoração do conjunto de recomendação.

Referências Bibliográficas

- BELÉM, F. et al. Associative tag recommendation exploiting multiple textual features. In: MA, W.-Y. et al. (Ed.). *SIGIR*. [S.l.]: ACM, 2011. p. 1033–1042. ISBN 978-1-4503-0757-4.
- BELÉM, F. M.; ALMEIDA, J. M.; GONÇALVES, M. A. A survey on tag recommendation methods. *JASIST*, v. 68, n. 4, p. 830–844, 2017.
- BELÉM, F. M. et al. Personalized and object-centered tag recommendation methods for Web 2.0 applications. *Information Processing & Management*, Elsevier, v. 50, n. 4, p. 524–553, 2014.
- BERLT, K. et al. A Graph Approach to Find Concepts in Short Textual Units. n. 16, 2015.
- FIGUEIREDO, F. et al. Evidence of Quality of Textual Features on the Web 2.0. In: *Proceedings of the 18th ACM Conference on Information and Knowledge Management*. New York, NY, USA: ACM, 2009. (CIKM '09), p. 909–918. ISBN 978-1-60558-512-3.
- GRAHAM, R.; CAVERLEE, J. Exploring Feedback Models in Interactive Tagging. In: *Web Intelligence*. [S.l.]: IEEE, 2008. p. 141–147.
- HONG, H.-K.; KIM, G.-W.; LEE, D.-H. Semantic tag recommendation based on associated words exploiting the interwiki links of Wikipedia. *Journal of Information Science*, SAGE Publications Sage UK: London, England, p. 0165551517693497, 2017.
- LIPCZAK, M. et al. Tag Sources for Recommendation in Collaborative Tagging Systems. In: EISTERLEHNER, F.; HOTH, A.; JÄSCHKE, R. (Ed.). *ECML PKDD Discovery Challenge 2009 (DC09)*. [S.l.: s.n.], 2009. (CEUR-WS.org, v. 497), p. 157–172. ISSN 1613-0073.
- LIPCZAK, M.; MILIOS, E. E. Efficient Tag Recommendation for Real-Life Data. *ACM TIST*, v. 3, n. 1, p. 2, 2011.
- MARTINS, E. F. et al. Measuring and Addressing the Impact of Cold Start on Associative Tag Recommenders. In: *Proceedings of the 19th Brazilian Symposium on Multimedia and the Web*. New York, NY, USA: ACM, 2013. (WebMedia '13), p. 325–332. ISBN 978-1-4503-2559-2.

- MARTINS, E. F. et al. On cold start for associative tag recommendation. *Journal of the Association for Information Science and Technology*, v. 67, n. 1, p. 83–105, 2016. ISSN 2330-1643. Disponível em: <<http://dx.doi.org/10.1002/asi.23353>>.
- MENEZES, G. V. et al. Demand-Driven Tag Recommendation. In: BALCÁZAR, J. L. et al. (Ed.). *ECML/PKDD (2)*. [S.l.]: Springer, 2010. (Lecture Notes in Computer Science, v. 6322), p. 402–417. ISBN 978-3-642-15882-7.
- NESS, S. R. et al. Improving Automatic Music Tag Annotation Using Stacked Generalization of Probabilistic SVM Outputs. In: *Proceedings of the 17th ACM International Conference on Multimedia*. New York, NY, USA: ACM, 2009. (MM '09), p. 705–708. ISBN 978-1-60558-608-3.
- OLIVEIRA, V. et al. Automatic Query Expansion Based on Tag Recommendation. In: *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. New York, NY, USA: ACM, 2012. (CIKM '12), p. 1985–1989. ISBN 978-1-4503-1156-4.
- PREISACH, C.; MARINHO, L. B.; SCHMIDT-THIEME, L. Semi-supervised Tag Recommendation - Using Untagged Resources to Mitigate Cold-Start Problems. In: ZAKI, M. J. et al. (Ed.). *PAKDD (1)*. [S.l.]: Springer, 2010. (Lecture Notes in Computer Science, v. 6118), p. 348–357. ISBN 978-3-642-13656-6.
- RAFAILIDIS, D. et al. Content-based Tag Propagation and Tensor Factorization for Personalized Item Recommendation Based on Social Tagging. *ACM Trans. Interact. Intell. Syst.*, ACM, New York, NY, USA, v. 3, n. 4, p. 26:1–26:27, jan. 2014. ISSN 2160-6455.
- Sandvine. 2016. Disponível em: <<https://www.sandvine.com/downloads/general/global-internet-phenomena/2016/global-internet-phenomena-report-latin-america-and-north-america.pdf>>. Acesso em: 23 ago. 2017.
- SCHEIN, A. I. et al. Methods and metrics for cold-start recommendations. In: *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM Press, 2002. p. 253–260.
- SIGURBJÖRNSSON, B.; Van Zwol, R. Flickr tag recommendation based on collective knowledge. In: ACM. *Proceedings of the 17th international conference on World Wide Web*. [S.l.], 2008. p. 327–336.
- WANG, M. et al. Assistive tagging: A survey of multimedia tagging with human-computer joint exploration. *ACM Comput. Surv.*, v. 44, n. 4, p. 25, 2012.
- Youtube. 2016. Disponível em: <<http://youtube-global.blogspot.com.br/2016/11/heres-to-eight-great-years.html>>. Acesso em: 14 mai. 2017.