1. Introdução

1.1. Motivação

Estudos indicam que o volume de páginas indexadas nos principais buscadores (**Google** e **Yahoo**) aproximam-se da casa dos 20 bilhões [2,3]. Desta forma, podemos induzir que o volume de conteúdo duplicado ou até mesmo de conteúdo não relevante também seja grande.

Assim sendo, não é difícil perceber que os usuários estão perdendo mais tempo para encontrar as informações que desejam e estão cada vez mais dependentes de mecanismos de busca. Ou seja, como os usuários não tem condições de absorver todo o conteúdo produzido, fica cada vez mais clara a necessidade de mecanismos que auxiliem os processos de busca e de tomada de decisão na Web.

Os sistemas de recomendação são sistemas responsáveis por facilitar o processo de tomada de decisão quando buscamos por informações na Web. Estes sistemas procuram organizar as informações baseados em lógicas previamente definidas que fazem sentido no contexto da informação procurada. Por exemplo, quando estamos navegando na Web em um Web site de compra de jogos e nos deparamos com uma lista dos dez jogos mais procurados pelos usuários, ou quando estamos numa locadora virtual e verificamos uma lista com os filmes mais vistos, estamos na verdade recebendo recomendações para facilitar nossas decisões.

Podemos afirmar, com base no exposto acima, que a aplicação de mecanismos de recomendação em Web sites de conteúdo noticioso, como o Portal G1, tendem a trazer bons resultados para os seus usuários do ponto de vista de informações relevantes.

1.2. Objetivo da dissertação

Diante do cenário descrito na Seção 1.1, propomos um sistema de recomendação de segundo nível que auxilie editores de notícias sugerindo conteúdos relacionados aos novos artigos produzidos com base apenas nas informações extraídas durante a confecção da nova notícia. O sistema é chamado de *segundo nível* pois apresenta recomendações aos editores de notícias para que, por sua vez, geram recomendações aos leitores das notícias.

Todavia, o processo de recuperação das matérias relacionadas baseado em informações existentes no texto e o modelo de ordenação dos resultados retornados nos permitem dizer que o trabalho se assemelha fortemente a um típico sistema de "query-by-document".

O ambiente de teste das técnicas desenvolvidas será o Portal G1. Atualmente a recomendação já é realizada pela equipe de editores do portal G1. Entretanto o procedimento de recomendação de notícias é manual e consiste em identificar palavras-chave relevantes ao texto, para em seguida consultar no sistema de busca interna ou mesmo no Google por textos que tratem do mesmo assunto. Depois de lidos os textos e entendido que alguma similaridade existe entre eles, o editor copia seus links e os cola em uma área especificada dentro do novo texto produzido. A esta área damos o nome de "saibamais". Desta forma, o processo de produção de conteúdo se torna mais lento porque o jornalista passa a dividir seu tempo entre a produção e a mineração de textos e recuperação de informação para realização de relacionamentos de conteúdos.

A expectativa do nosso sistema é prover um mecanismo automático de recomendação, que forneça ao jornalista uma lista de possíveis matérias relacionadas, de modo a minimizar o tempo gasto na produção de notícias da redação. O objetivo intermediário é promover uma melhora no conteúdo relacionado utilizando diferentes critérios para geração das recomendações.

1.3. Organização da dissertação

O restante deste documento está organizado em cinco capítulos, da seguinte forma. O Capítulo 2 apresenta o estado da arte em sistemas de recomendação e as ferramentas e *frameworks* existentes comparados a ferramenta proposta. O Capítulo 3 descreve a ferramenta desenvolvida para apoiar a recomendação de

conteúdo na produção de matérias. O Capítulo 4 apresenta os resultados obtidos com a experimentação da ferramenta contra um corpus real. O Capítulo 5 apresenta as conclusões e trabalhos futuros.