1 Introdução

A popularização dos computadores possibilitou o armazenamento cada vez maior de conteúdos digitais, entre esses, o formato textual como em livros, documentos, e-mails, redes sociais e páginas web. A produção de textos gera fontes de informações em volumes crescentes que podem superar a capacidade humana de analisá-los manualmente. Essa dificuldade incentiva a pesquisa de ferramentas automáticas para manipulação de dados não estruturados. Assim, os processos de extração automática de conhecimento em coleções textuais são essenciais, e ao mesmo tempo, constituem um desafio, devido às características de documentos textuais como o formato não estruturado e trechos com diferentes níveis e importância, desde informações essenciais até textos pouco informativos e em alguns casos até irrelevantes.

Além dos tipos de informações mais comuns que são armazenados no formato textual, como e-mails, relatórios, artigos e postagens em redes sociais, têm-se também o armazenamento das atas de reuniões, as quais permitem às organizações a documentação oficial de reuniões em arquivos digitais, facilitando a confecção, compartilhamento e consulta às decisões tomadas. Reuniões são tarefas presentes em atividades corporativas, ambientes de gestão e organizações de um modo geral, onde discute-se problemas, soluções, propostas, planos, questionamentos, alterações de projetos e frequentemente são tomadas decisões importantes. A comunicação entre os membros da reunião é feita de fo majoritariamente verbal. Para que seu conteúdo possa ser registrado e externalizado, adota-se a prática de registrar seu conteúdo em documentos, ~~os quais chamamos de atas de reunião.~~

~~Porém, armazenar e recuperar atas em formato textual oferece desafios.~~ As atas de reunião possuem características particulares, ~~fre~~ mente apresentam um texto com poucas quebras de parágrafo e sem marcações de estrutura, como capítulos, seções ou quaisquer indicações sobre o tema do texto. Devido a fatores como a não estruturação e volume dos textos, a localização de um assunto em uma coleção de atas é uma tarefa custosa, especialmente considerando o seu crescimento em uma instituição. As organizações costumam manter seus documentos eletrônicos organizados em pastas e nomeá-los com informações básicas sobre a reunião a que se refere como a data e alguma referência cronológica, por exemplo "37ª Reunião Ordinária do Conselho ...". Essa forma de organização facilita a localização dos arquivos com ferramentas que fazem buscas pelo nome dos arquivos e pastas. Contudo, essa prática costuma ser insuficiente, pois uma busca pelo conteúdo dos textos usa-se ferramentas computacionais baseadas em localização de palavras-chave que além de encontrar ocorrências das palavras podem oferecer recursos como operadores *and*

e *not* ou ainda suporte às expressões regulares. Esse recurso, conhecido como *grepping*[1], traz resultados satisfatórios em muitos casos. Por outro lado, o *grepping* traz algumas desvantagens como: 1) transfere o podem ser a tarefa ao usuário; 2) buscas em grandes coleções de documentos ~~tornam-se mais~~ lentas; 3) não há suporte a padrões mais flexíveis como a proximidade entre as palavras ou palavras que estejam na mesma sentença; 4) o retorno ao usuário são os documento integrais, o que pode exigir uma se precisa colocar referências de trabalhos que têm feito isso dentro de um documento para encontrar o trecho desejado.

Para superar essas limitações tem sido utilizadas técnicas de ap de máquina por meio de diversas abordagens. Por exemplo, elas vêm sendo empregadas na organização, gerenciamento, recuperação de informação e extração de conhecimento, como a ext categorização de automática de documentos. Essas técnicas perm a por informações em atas de reunião.

esses conteúdos estão sem ligação. Acho que o paragrafo de baixo é mais ligado à motivação da utilidade das atas. Me parece que ele encaixa melhor como 3o parágrafo.

nselho de um programa de pós-graduação de uma universidade, são d , quais são os critérios para credenciamento e permanência de docen longo do tempo, esse tema pode ser discutido e mencionado diver critérios inclusive passar por significativas alterações, devido a diversos fatores. O coordenador do programa pode desejar recuperar qual foi a decisão mais recente, para poder aplicar os critérios a um potencial novo membro do programa, ou os membros do conselho podem desejar rever o histórico de tudo o que já foi discutido/decidido sobre o tema, para poder propor alterações nas regras, de forma mais adequada.

Uma vez que a ata registra a sucessão de assuntos discutidos na reunião, um sistema de recuperação de informação idealmente deve retornar ao usuário apenas o trecho que trate do assunto pesquisado ao invés do documento inteiro. Assim, cada trecho com um assunto predominate pode ser considerado um subdocumento. Portanto, em primeiro lugar, há a necessidade de descobrir onde há mudanças de assunto no texto, ~~que pode ser atendida com técnicas de segmentação automá~~. Técnicas de segmentação automática de textos (segmentação textual) podem ser aplicadas com esse propósito.

A tarefa de segmentação automática de textos, ou segmentação textual consiste em dividir um texto em partes que contenham um significado relativamente independente. Em outras palavras, é identificar as posições nas quais há uma mudança significativa de assunto. É útil em aplicações que trabalham com textos sem indicações de quebras de assunto, ou seja, não apresentam seções ou capítulos, como transcrições automáticas de áudio, víd grandes documentos que contêm vários assuntos como atas de reunião e notícias. Pod usada para melhorar o acesso a informação solicitada por meio de uma consulta, onde é possível oferecer porções menores de texto mais relevantes ao invés de exibir um documento grande que pode conter informações menos pertinentes. A navegação pelo documento pode ser aprimorada, em especial na utilização por usuários com deficiência visual, os quais utilizam sintetizadores de texto como ferramenta de acessibilidade. Além disso, encontrar

colocar uma referencia disso

[1] O nome *grepping* é uma referência ao comando `grep` do Unix

não entendi o sentido de mencionar isso aqui

pontos onde o texto muda de assunto, pode ser útil como etapa de pré-processamento em aplicações voltadas ao entendimento do texto, principalmente em textos longos (CHOI, 2000).

Assim, nesse contexto, este trabalho propõe a investigação do uso de mineração de texto e as técnicas que constituem o estado da arte na área para o desenvolvimento de uma ferramenta para extração automática de históricos de decisão em atas de reuniões.