Foogle

Grupo: Luiz Filipe Martins Ramos, Marcio Valença Ramos e Felipe Romero Pereira.

Data: 10/10/13

Descrição do projeto: Será desenvolvido um website capaz de realizar pesquisas de palavras e frases dentro do Facebook do usuário. Essas pesquisas podem ser tanto no chat, quanto no newsfeed, nos grupos e nas fotos do usuário. Para essa pesquisa ser eficiente será necessário indexar parte do banco de dados do Facebook a partir das palavras e propor alguns algoritmos de Inteligência Artificial para ranquear os resultados obtidos.

Requisitos (em ordem de prioridade):

- Pesquisa simples por palavra única no chat, nos posts e nas fotos.
- Pesquisa por múltiplas palavras.
- Pesquisa por trechos (por exemplo, “Game of Thrones”).
- Redução para palavras similares (exemplo, “você” = “voce” = “vc”).
- Autocomplete de pesquisas frequentes.

Materiais utilizados: O banco de dados do site ficará a principio, armazenado nos servidores do Google. Será utilizado o Google App Engine em Python (cujo banco de dados é NDB). A interface do site será feita em HTML5 com CSS e Javascript. O sistema de versionamento de código escolhido foi o Git por ser o mais familiar aos três integrantes da equipe. As consultas ao banco de dados do Facebook serão feitas em FQL (Facebook Query Language) a partir da API do próprio Facebook.

Pesquisas iniciais: Foram realizadas algumas pesquisas iniciais sobre o funcionamento do banco de dados do Facebook para que seja possível realizar as queries necessárias. Observou-se que todos os objetos do site possuem identificadores numéricos únicos. Os usuários, comentários e mensagens de chat são identificados por um long, enquanto que os posts são identificados por dois longs. Além disso, observou-se que no banco de dados do site, a nomenclatura utilizada é a seguinte: comentários estão na tabela *comment*, posts estão na tabela *stream* e mensagens de chat estão na tabela *message*. O diagrama EER abaixo foi criado para se facilitar a visualização e auxiliar na construção de queries.

Queries uteis: Foram geradas várias queries em FQL (muito similar ao SQL) que serão uteis no desenvolvimento do projeto.

- Pegar as mensagens de todos os chats do usuário:

```
SELECT body, message_id FROM message WHERE thread_id IN
    (SELECT thread_id FROM thread WHERE folder_id=0)
    ORDER BY created_time DESC
```

- Pegar todos os posts que estão na linha do tempo do usuário:

SELECT post_id, message FROM stream WHERE source_id=me() LIMIT 1000

- Pegar todos os posts de todos os grupos dos quais o usuário pertence:

SELECT post_id, message FROM stream WHERE source_id IN
(SELECT gid FROM group_member WHERE uid=me())

- Pegar todos os comentários de todos os posts de todos os grupos dos quais o usuário pertence:

SELECT text FROM comment WHERE post_id IN
(SELECT post_id FROM stream WHERE source_id IN
(SELECT gid FROM group_member WHERE uid=me()))

- Pegar todos os posts do newsfeed do usuário:

SELECT post_id, message FROM stream WHERE filter_key IN
(SELECT filter_key FROM stream_filter WHERE type = 'newsfeed' AND uid=me())
AND is_hidden=0

O diagrama EER da figura 1 foi desenvolvido para facilitar a compreensão das queries e desenvolvimento de outras. Ele compreende a parte do banco de dados do Facebook relevante para o projeto.

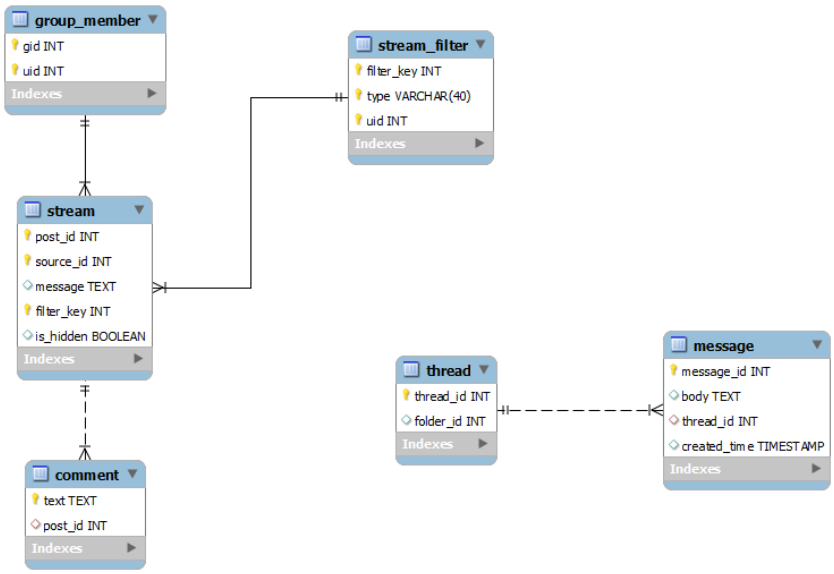

Figura 1: EER da parte relevante do banco do facebook.

Problemas encontrados: Nessa parte inicial do projeto foram identificados alguns problemas. Primeiramente o Facebook só retorna no máximo 30 mensagens do chat de uma pessoa de cada vez. Então para recuperar uma quantidade considerável será necessário realizar muitas queries, deixando o resultado muito lento. Outro problema encontrado é o tamanho do banco

de dados gratuito disponibilizado pelo Google (apenas 1Gb) que é muito pequeno para a quantidade de dados queremos armazenar a princípio.

Para se ter uma melhor noção da ordem de grandeza da quantidade de dados, foram realizados vários testes, usando Javascript para fazer as queries e analisar o resultado.

	Teste #1	**Teste #2**	**Teste #3**
Número de mensagens analisadas	659	674	858
Número total de palavras	4825	4546	4432
Número de palavras distintas	1578	1535	1462
Palavra mais frequente	“e” – 163 vezes	“o” – 160 vezes	“o” – 128 vezes

Tabela 1: Resultado da análise estatística sobre mensagens no chat

Assim, em um conjunto de aproximadamente 4600 palavras, temos em média 6.4 palavras por mensagem e 3 instâncias de uma mesma palavra.

Além disso, a partir deste teste notou-se a necessidade de filtrar palavras comuns que não agregam valor à busca, como artigos por exemplo.