Organização e Recuperação da Informação

Wendel Melo
Faculdade de Computação
Universidade Federal de Uberlândia

Recuperação da Informação

Adaptado do Material da Profª Vanessa Braganholo - IC/UFF

Recuperação de Informação

- Recuperação de Informação (RI) se remete a, dada uma **base de documentos**, ser capaz de apontar um subconjunto destes que atendam à necessidade de informação do usuário;

- Idealmente, os documentos devem ser apresentados segundo um ***ranking*** onde os supostamente mais relevantes vêm antes dos menos relevantes;

- Todavia, a **relevância** é um conceito subjetivo que pode depender de diversos fatores externos como localização, instante de tempo, dispositivo, preferências pessoais, nível de cultura, etc;

Recuperação de Dados

- Tarefas determinísticas e precisas;
- Respostas devem ser corretas;
- Sistemas não visam incorporar o significado do que está sendo buscado.
- Ex:
  - Obter lista alunos de SI com CRA maior que 80;
  - Busca por documentos com a palavra Brasil.

X

Recuperação de Informação

- Tarefas imprecisas;
- Pequenos erros são tolerados;
- Normalmente não há o conceito de resposta 100 % correta;
- Sistemas se preocupam com o significado do que está sendo buscado;
- Ex: busca por **bons** documentos sobre o Brasil.

Recuperação de Dados

- A informação pode estar bem estruturada como em banco de dados, o que permite mecanismos de recuperação elaborados como consultas SQL;
- Pode se mostrar limitada quando for preciso trazer informações sobre um determinado assunto.

X

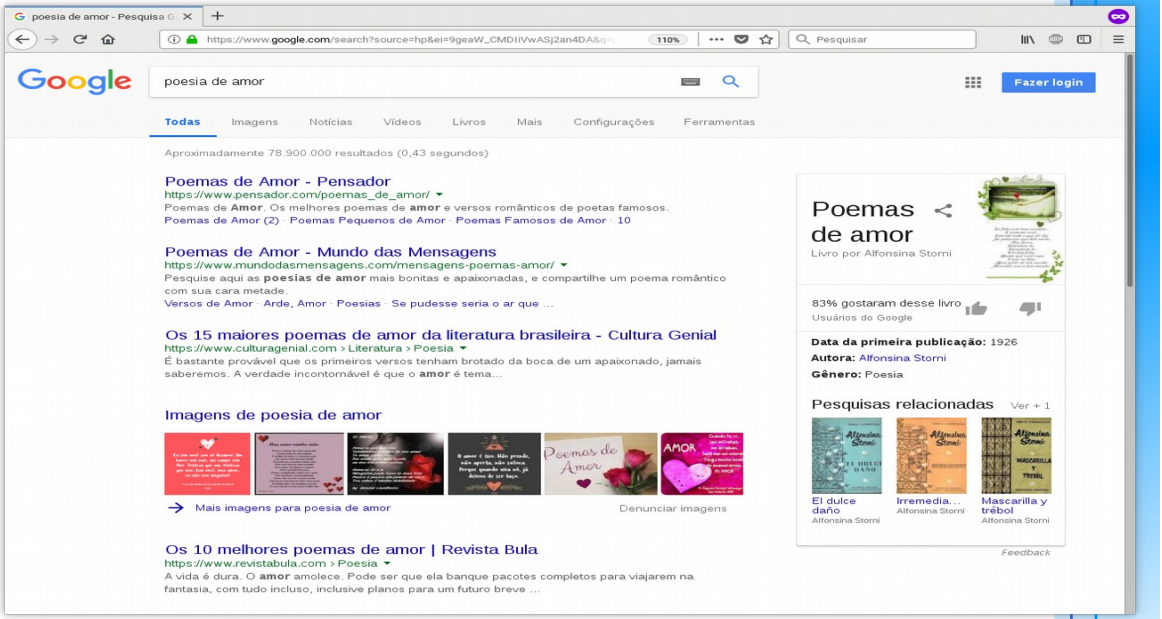
Recuperação de Informação

- Frequentemente lida com textos em linguagem natural;
- Documentos da base costumam não ser bem estruturados e podem ser semanticamente ambíguos;
- Pode organizar e consultar acervos de documentos.
- Em geral, não há suporte para consultas bem elaboradas como em SQL.

Recuperação de Informação

- Exemplo de sistema de RI: buscador de internet. Base de documentos: conteúdo da WEB

- Uma busca por “poesia de amor” no Google retornou dezenas de milhões de resultados (já ranqueados) em menos de meio segundo!



Recuperação de Informação

- Nesse curso, estudaremos técnicas da área de recuperação da informação, que dão a base para que seja possível ter um sistema como o buscador Google em funcionamento;
- Veremos que área de RI é altamente empírica, o que abre espaço para criatividade de técnicos e acadêmicos;
- O sucesso de sistemas como Google se deve, em parte, a engenhosas ideias para melhorar eficiência e eficácia sobre uma base de dados gigantesca;
- Entretanto, muitos sistemas de RI podem ter uma base não tão grande para pesquisar.

Recuperação de Informação

- A **base de documentos** sobre a qual um sistema de RI atua depende do contexto e pode ser composta de:
  - Livros;
  - Documentos;
  - Imagens;
  - Áudios;
  - Vídeos;
  - Catálogos;
  - Prontuários de pacientes;
  - Páginas da internet;
  - Normas
  - Notícias
  - Registros em geral:
    - Estruturados;
    - Semiestruturados;
    - Não estruturados

Visões de Recuperação de Informação

- A área de RI possui duas visões complementares:
- **Centrada no computador:** consiste principalmente na construção de estruturas de dados eficientes, no processamento de consultas com alto desempenho e desenvolvimento de bons modelos e algoritmos de ranqueamento.
- **Centrada no usuário:** estuda do comportamento do usuário, o entendimento de suas principais necessidades e como estas afetam a organização e a operação do sistema de recuperação.
- A visão centrada no computador é o foco da disciplina e historicamente tem recebido maior atenção (mas a atenção pela visão centrada no usuário também vem crescendo).

Histórico da área de RI

- Os primeiros sistemas computacionais de RI surgiram para automatizar acesso a informação em bibliotecas na década de 1960;
- Até o início dos anos 1990, as aplicações principais da área ainda eram catálogos de bibliotecas, jornais, revistas e enciclopédias eletrônicas, além de bases de dados de empresas;
- Até então, RI era uma área periférica dentro da computação, contando com a atuação de poucos pesquisadores e técnicos.

Histórico da área de RI

- No final dos anos 90, uma mudança brusca trouxe RI para o primeiro plano: a popularização da Web.

- Junto com a Web, surgiram novos desafios, por exemplo:

1) Base de dados bastante distribuída: é preciso coletar os documentos para um repositório central;

2) Base de dados muito extensa: é fundamental um bom ranqueamento;

3) Grande número de usos simultâneos: problemas de escalabilidade e desempenho.

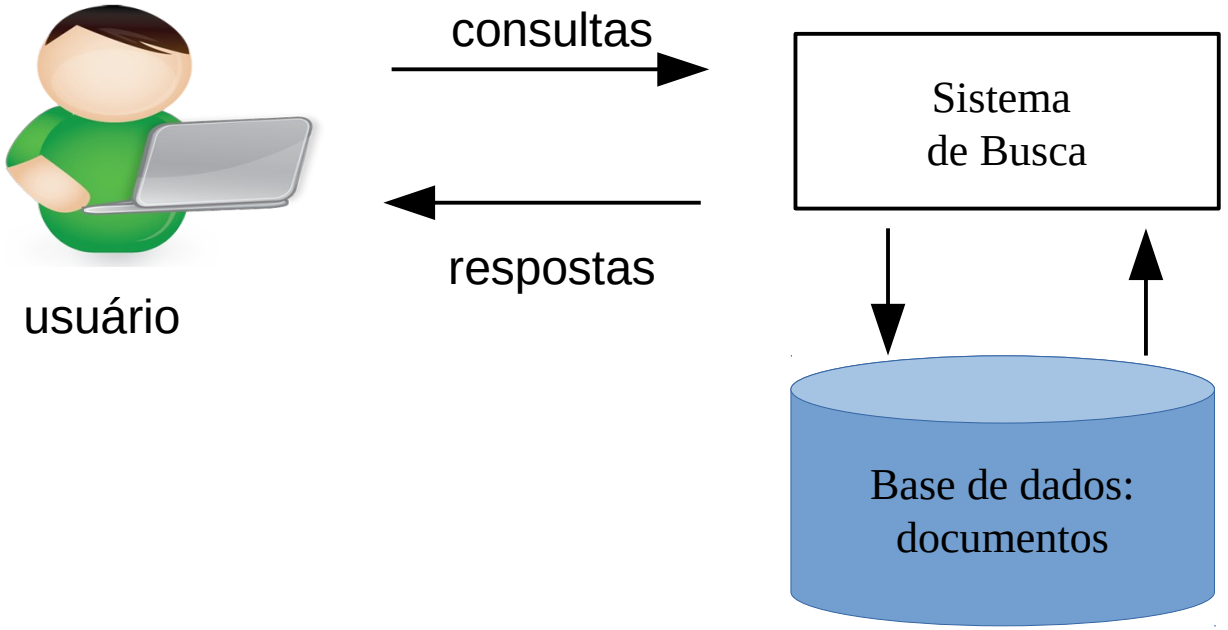
Tipos de problemas de RI

- Os principais tipos de problemas na área de RI são:
  - Busca
  - Filtragem
  - Classificação

Busca

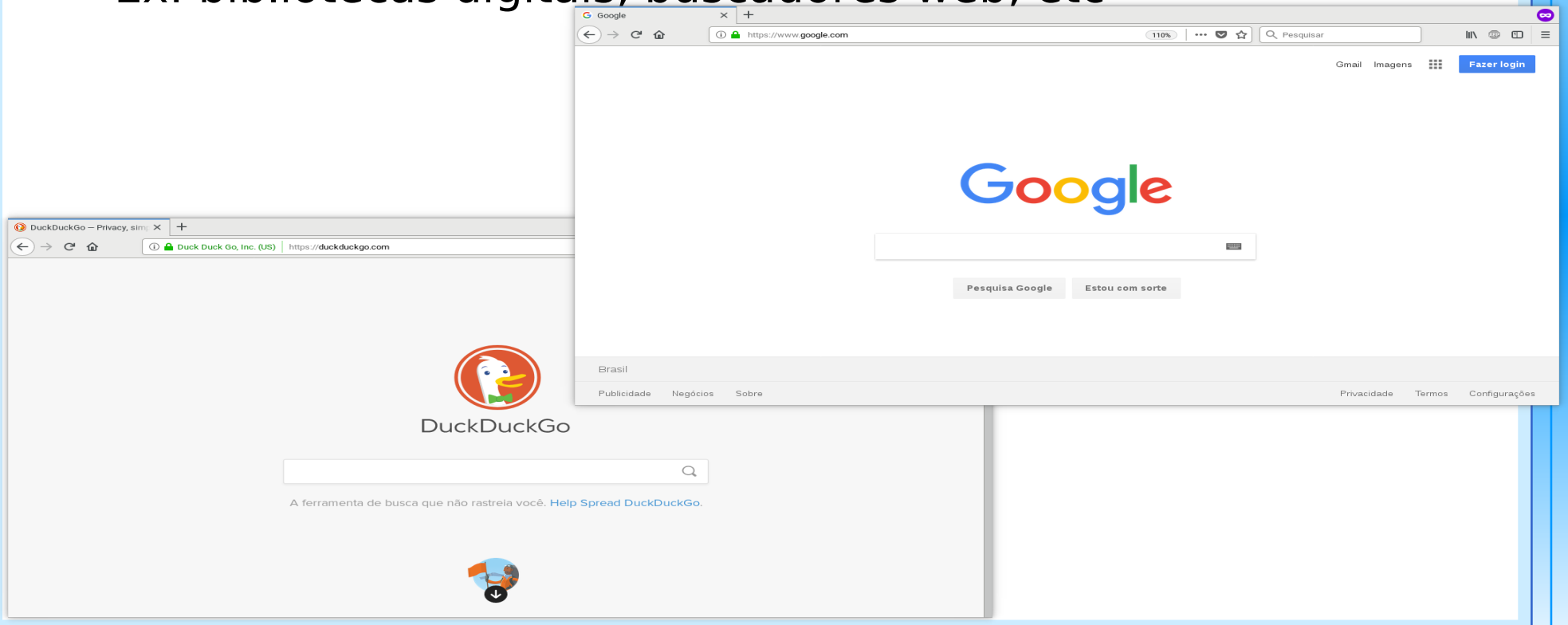
- **Base de dados:** documentos.
- **Entrada:** consultas dos usuários.
- **Objetivo:** retornar os documentos que melhor atendem às consultas.
- Usuários apresentam uma consulta e o sistema busca respostas em uma base de dados pré-existente.
- Tipo mais comum

Sistemas de Busca - Funcionamento



Busca

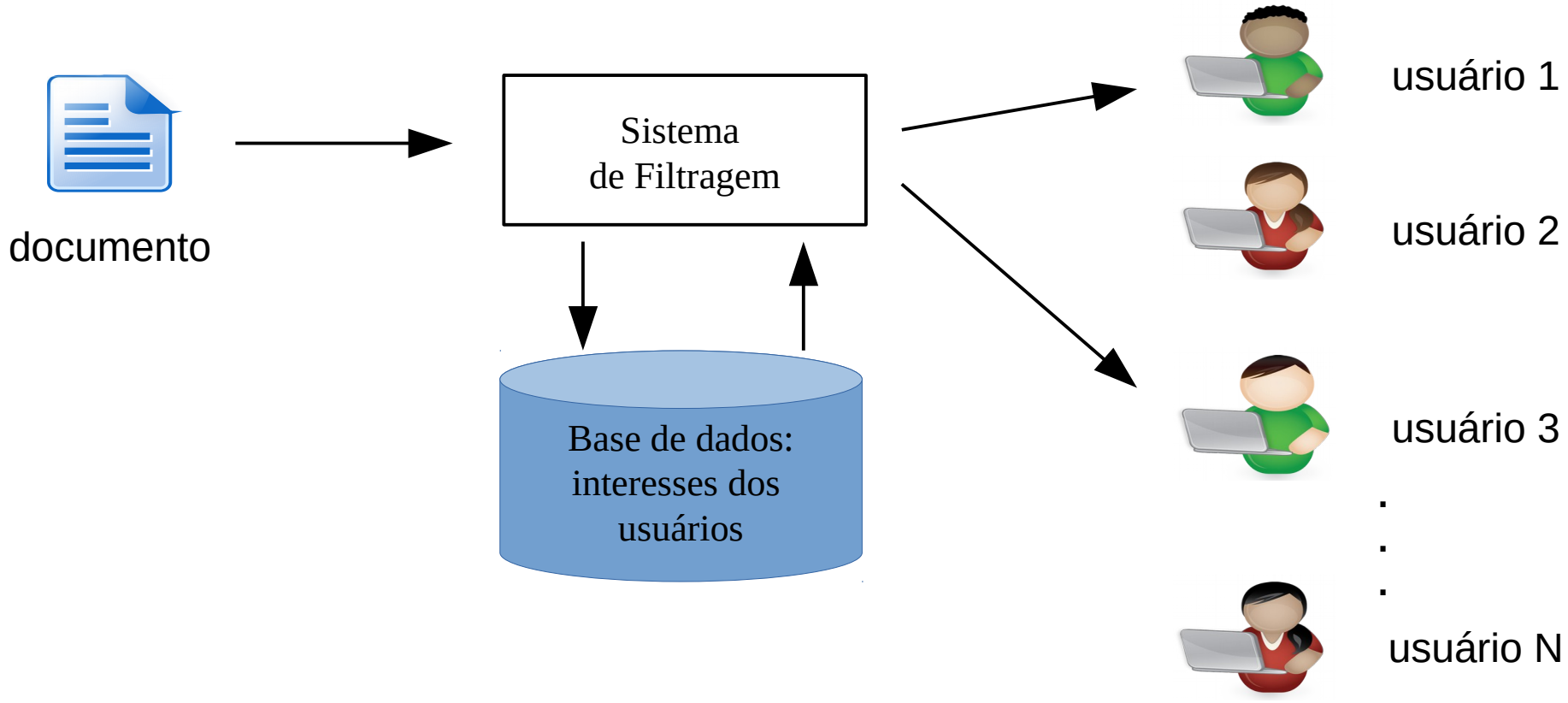
- Ex: bibliotecas digitais, buscadores web, etc



Filtragem

- **Base de dados:** lista de interesses de cada usuário.
- **Entrada:** documentos.
- **Objetivo:** identificar os usuários que se interessam pelos documentos.
- Inverso do problema de busca;
- Aqui, os interesses do usuário estão pré-cadastrados e os documentos vão chegando ao sistema dinamicamente, que então identifica possíveis interessados nos documentos.

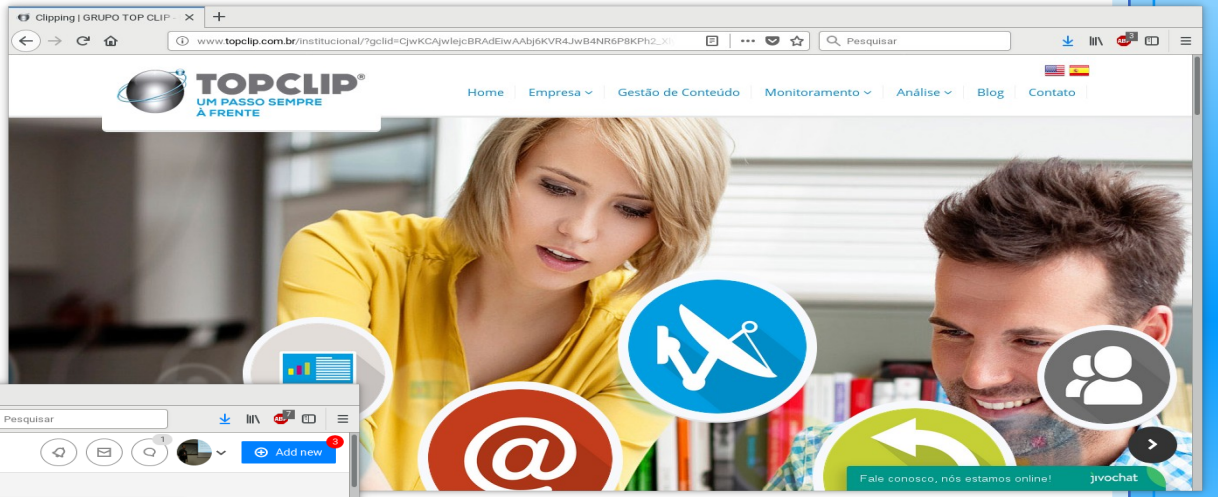
Sistemas de Filtragem - Funcionamento

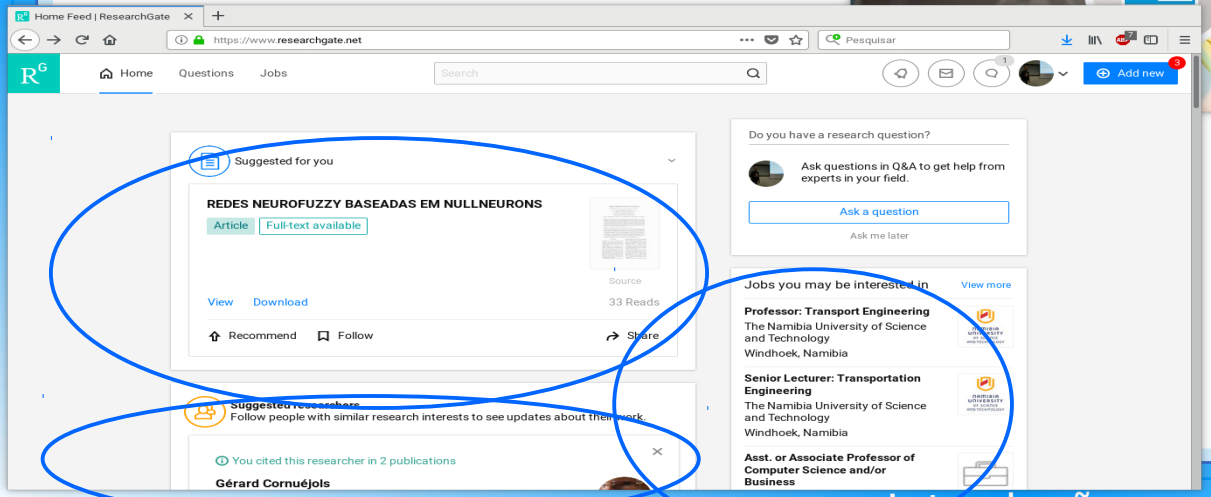


Filtragem

www.topclip.com.br: sistema de monitoramento de mídias

- Usado em sites de notícias, controle de correspondência, sistemas de publicações, etc.



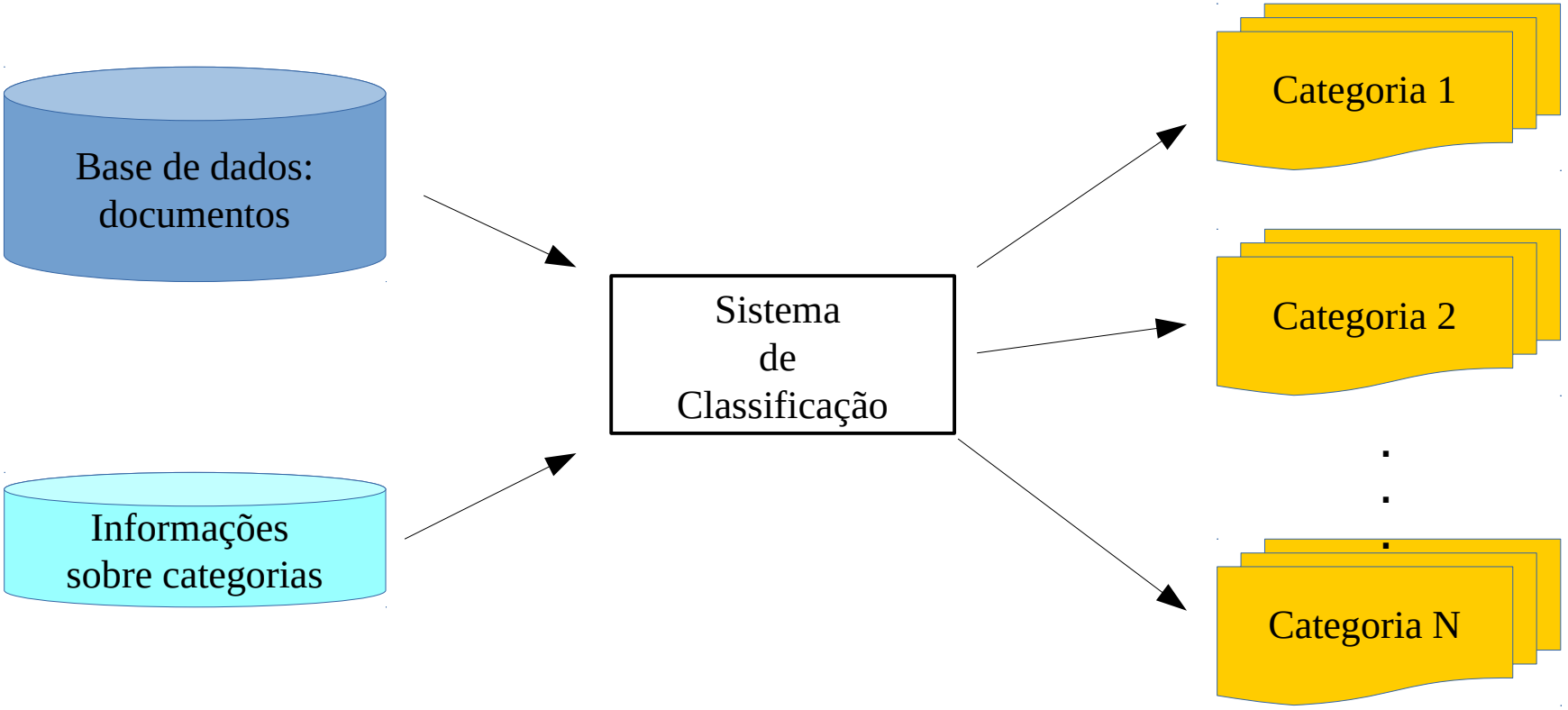


www.researchgate.net: rede social para pesquisadores com sugestões de artigos, empregos e autores.

Classificação

- **Base de dados:** documentos e descrição de categorias de documentos.
- **Objetivo:** enquadrar os documentos nas categorias adequadas.
- Quando as categorias não são conhecidas, o problema é conhecido como problema de agrupamento (clustering)

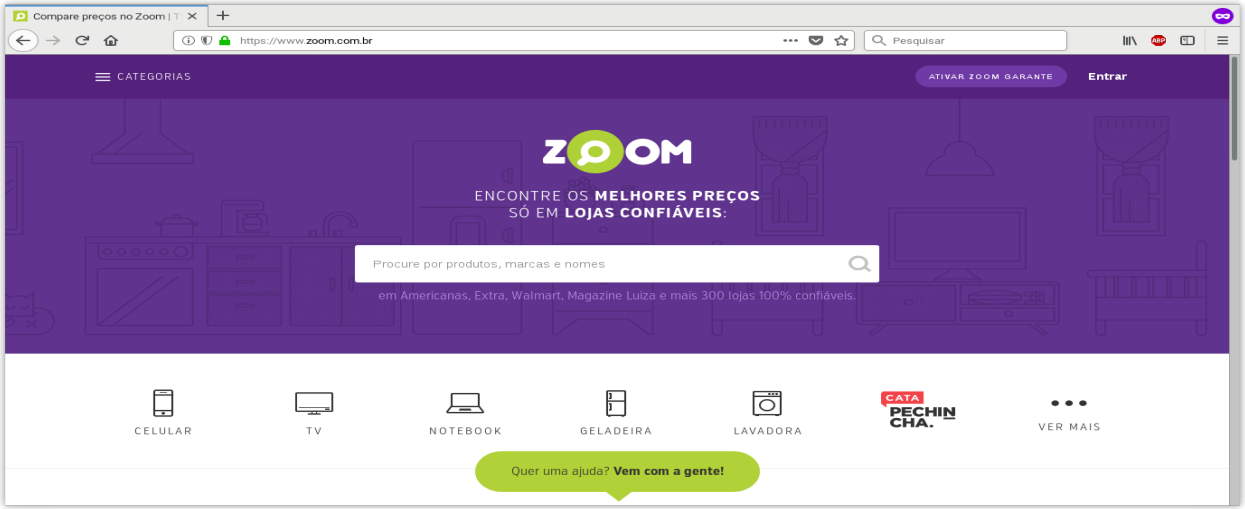
Sistemas de Classificação - Funcionamento



Classificação

- Ex: sistemas de monitoramento de lojas eletrônicas

www.zoom.com.br



Recuperação de Informação

- O foco da disciplina são os sistemas de busca;
- Em alguns casos, sistemas podem mesclar diferentes problemas de RI;
- Quando o usuário fornece termos para uma pesquisa, dizemos que o mesmo está realizando uma **busca**;
- Quando o usuário clica em links para navegar em categorias, dizemos que está realizando uma **navegação**;
- Alguns sistemas de RI podem mesclar busca e navegação.