



INSTITUTO FEDERAL  
MINAS GERAIS

# RECUPERAÇÃO DE INFORMAÇÃO

Profa. Patrícia Proença

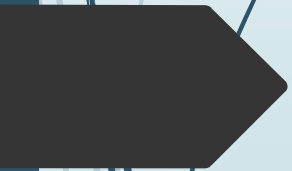
[patricia.proenca@ifmg.edu.br](mailto:patricia.proenca@ifmg.edu.br)



# ATENÇÃO!!!

- ↴ O material a seguir é uma videoaula apresentada pela professora PATRÍCIA APARECIDA PROENÇA AVILA, como material pedagógico do IFMG, dentro de suas atividades curriculares ofertadas em ambiente virtual de aprendizagem. Seu uso, cópia e ou divulgação em parte ou no todo, por quaisquer meios existentes ou que vierem a ser desenvolvidos, somente poderá ser feito, mediante autorização expressa deste docente e do IFMG. Caso contrário, estarão sujeitos às penalidades legais vigentes”.
- ↴ Conforme Art. 2º§1º da Nota Técnica nº 1/2020/PROEN/Reitoria/IFMG (SEI 0605498, Processo nº 23208.002340/2020-04

# Realimentação de Relevância e Expansão de Consultas





# Roteiro

- Introdução;
  - Informações explícitas;
  - Informações implícitas;
- Métodos de realimentação explícitos (realimentação de relevância):
  - Avaliação da realimentação;
  - Cliques;
- Métodos de realimentação implícitos (expansão de consultas):
  - Local (Análise de contexto);
  - Global (Tesauro de similaridade).



# Introdução

- Sem um conhecimento detalhado da coleção, a maioria dos usuários acha difícil formular consultas bem projetadas para fins de recuperação;
- Exemplo: usuários de sistemas de RI muitas vezes precisam reformular suas consultas para obter os resultados que interessam
  - A primeira consulta deve ser tratada como uma tentativa inicial de recuperar informações relevantes!
  - Formulações melhores da consulta podem ser escritas para recuperar mais documentos úteis.




# Introdução

- Como melhorar a formulação da consulta inicial utilizando a informação que está relacionada com a intenção “por trás” da consulta?
  - 1) Realimentação de relevância (realimentação explícita) – quando o usuário fornece explicitamente informações sobre os documentos relevantes para uma consulta;
  - 2) Expansão de consultas (realimentação implícita) – quando informações relacionadas à consulta são utilizadas implicitamente pelo sistema.



# Métodos de realimentação

- **Definição:** *A realimentação de relevância refere-se a um ciclo de realimentação em que documentos que são conhecidamente relevantes para a consulta  $q$  em questão são usados para transformá-la em uma consulta modificada  $q_m$ .*
- A expectativa é que a consulta  $q_m$  retornará um maior número de documentos relevantes para  $q$ .



# Métodos de realimentação

## Problemas

- Obter informações sobre a relevância dos documentos em relação a consulta:
  - É caro;
  - Exige a interferência direta do usuário;
- Exemplo: um sistema de RI poderia perguntar aos usuários se os 10 primeiros resultados para uma determinada consulta são de fato relevantes. Será que os usuários estão dispostos a fornecer essa informação?





# Métodos de realimentação

## Possível solução

- Em vez de pedir aos usuários que marquem os documentos relevantes, poderíamos analisar documentos que:
  - Eles tenham clicado;
  - Ou observar os termos pertencentes aos documentos do topo do conjunto dos resultados.
- Em ambos os casos, se supusermos que a informação recolhida está relacionada à consulta original, esperamos que o ciclo de realimentação produza resultados de melhor qualidade!



# Ciclo de Realimentação

## Etapas

- 1) Determinar a informação de realimentação que está relacionada, ou que se espera que esteja relacionada à consulta original;
- 2) Determinar como transformar a consulta q de modo a utilizar essa informação de forma eficaz.

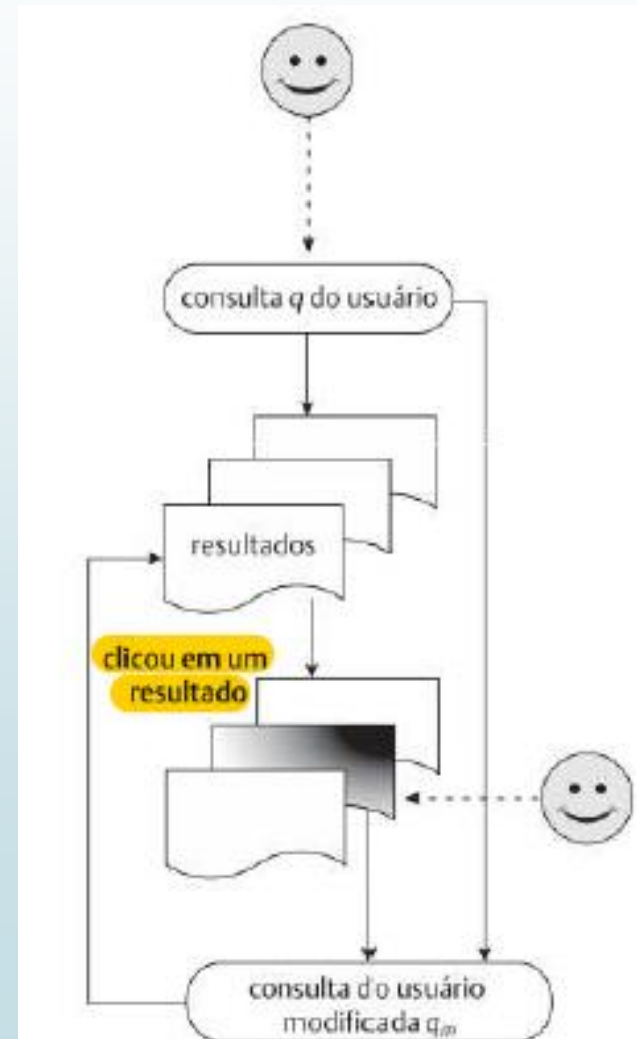
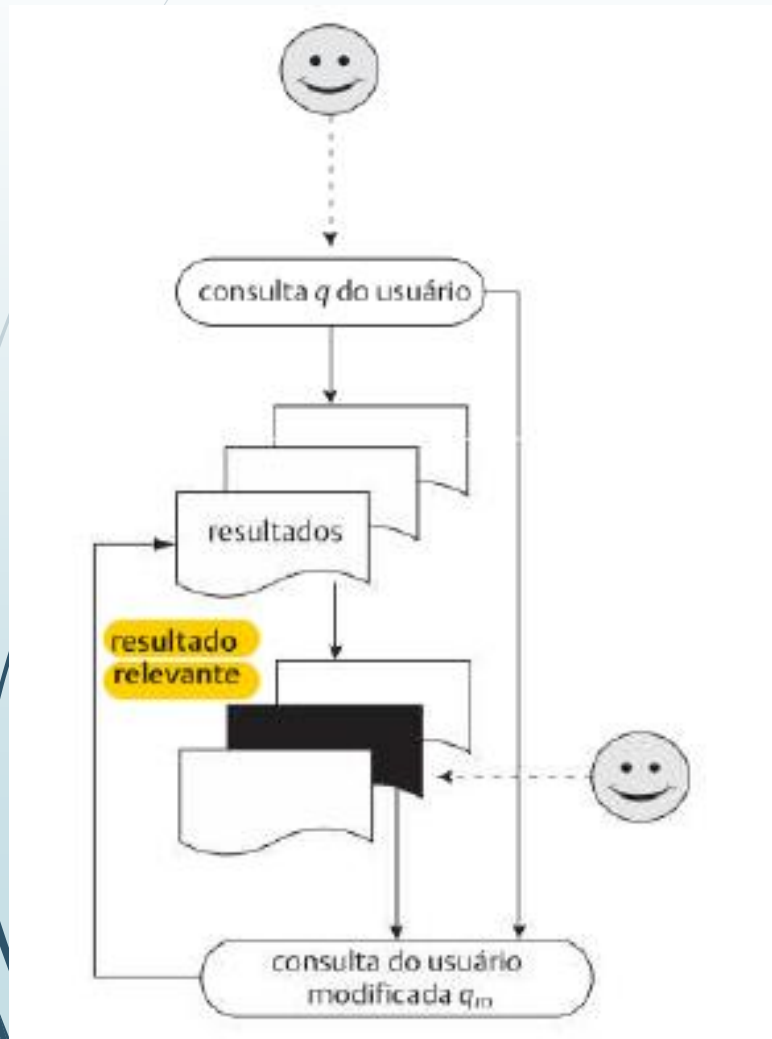


# Ciclo de Realimentação

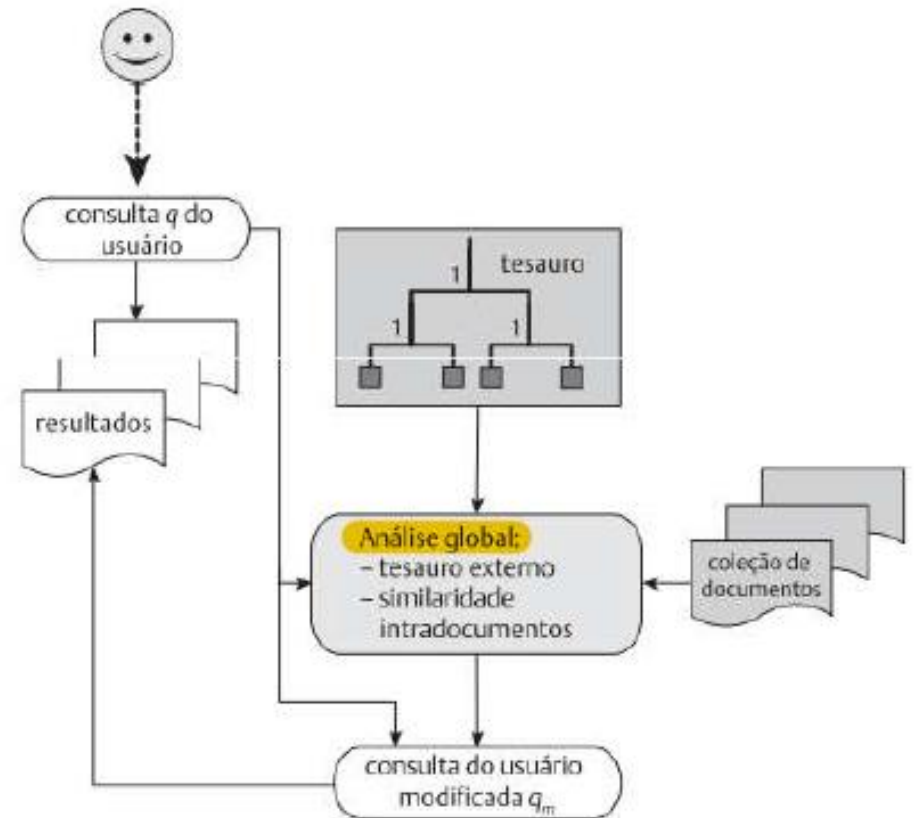
## Etapas

- A etapa 1 pode ser realizada de duas formas distintas:
- a) Obter explicitamente a informação de realimentação a partir dos usuários;
- b) Obter implicitamente a informação de realimentação a partir dos resultados da consulta ou de fontes externas;

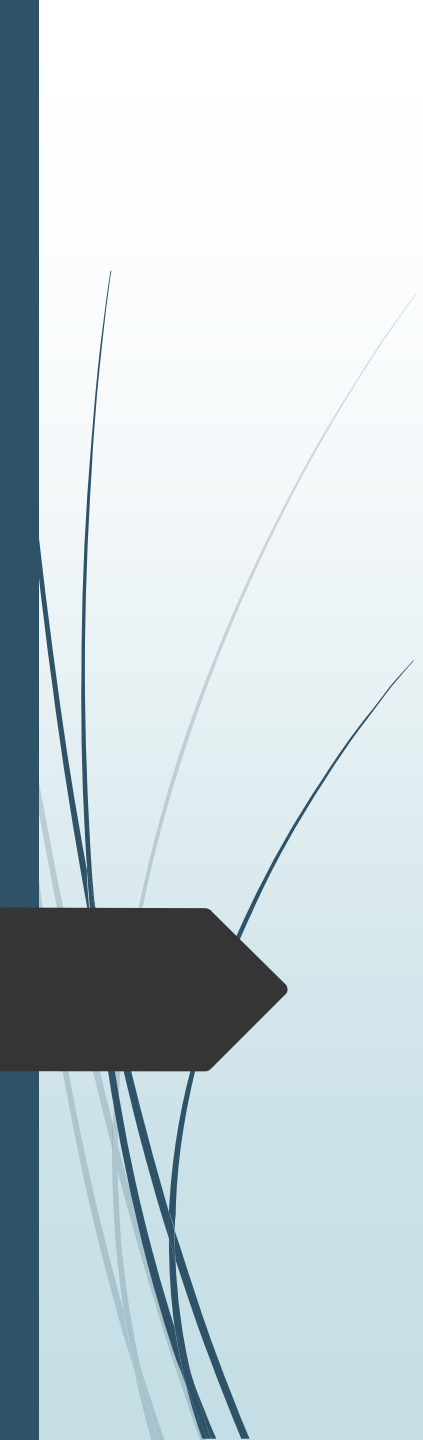
# Informações explícitas de realimentação



# Informações implícitas de realimentação



Ambos não tem a participação do usuário.



# Métodos de realimentação de relevância explícitos



# Realimentação explícita

## Cliques

- Usuários de máquinas de busca na Web não só inspecionam os resultados de suas consultas, como também clicam sobre eles;
- Os cliques podem ser coletados em grandes números, sem interferir nas ações dos usuários;
- Pergunta: os dados de cliques podem ser usados para decidir sobre a relevância do resultado para consultas futuras?



# Realimentação explícita

## Comportamento do usuário

- Experimentos realizados com um grupo de 29 indivíduos;
- Resultados:
  - Usuários escaneiam os resultados da consulta de cima para baixo;
  - Inspeccionam o primeiro e o segundo resultados de imediato;
  - Tendem a escanear em detalhe as primeiras cinco ou seis respostas que aparecem na área visível da tela;





# Realimentação explícita

## Comportamento do usuário

- Resultados:
  - 60-70% das tarefas, os usuários têm uma fixação no primeiro ou no segundo resultado – para o quarto resultado a frequência cai pela metade;
  - Usuários inspecionam as duas primeiras respostas quase que igualmente, mas eles clicam quase três vezes mais no primeiro resultado;
  - Indicativo de que o usuário tende a confiar na máquina de busca!



# Realimentação explícita

## Cliques

- Participantes recebem dois conjuntos distintos de resultados:
  - O ranking normal retornado pela máquina de busca;
  - Um ranking modificado, no qual os dois melhores resultados têm a sua posição trocada.
- O que acontece?
  - Usuários clicam quase três vezes mais no primeiro resultado do que no segundo!
  - A **posição do resultado tem uma grande influência** na decisão do usuário.



# Realimentação explícita

## Cliques

- Interpretar cliques como um **indicativo direto** de relevância não é uma boa abordagem...
  - Ideia: interpretar cliques como métricas de PREFERÊNCIA do usuário;
  - Exemplo: Se você olha para o conjunto de um resultado e decide ignorá-lo e clicar em um resultado mais abaixo no ranking, é apropriado dizer que este usuário prefere o resultado clicado ao mostrado mais acima do ranking.

# Cliques dentro de uma mesma consulta

- Considere o exemplo sobre o comportamento de um usuário que clique sobre as respostas de um conjunto de resultados:

$r_1 \quad r_2 \quad \sqrt{r_3} \quad r_4 \quad \sqrt{r_5} \quad r_6 \quad r_7 \quad r_8 \quad r_9 \quad \sqrt{r_{10}}$

- O usuário clicou nos resultados  $r_3$ ,  $r_5$  e  $r_{10}$ , mas não nos permite tirar conclusões definitivas sobre a relevância da consulta.
  - No entanto fornece informações sobre preferências desse usuário.



# Realimentação de relevância explícita

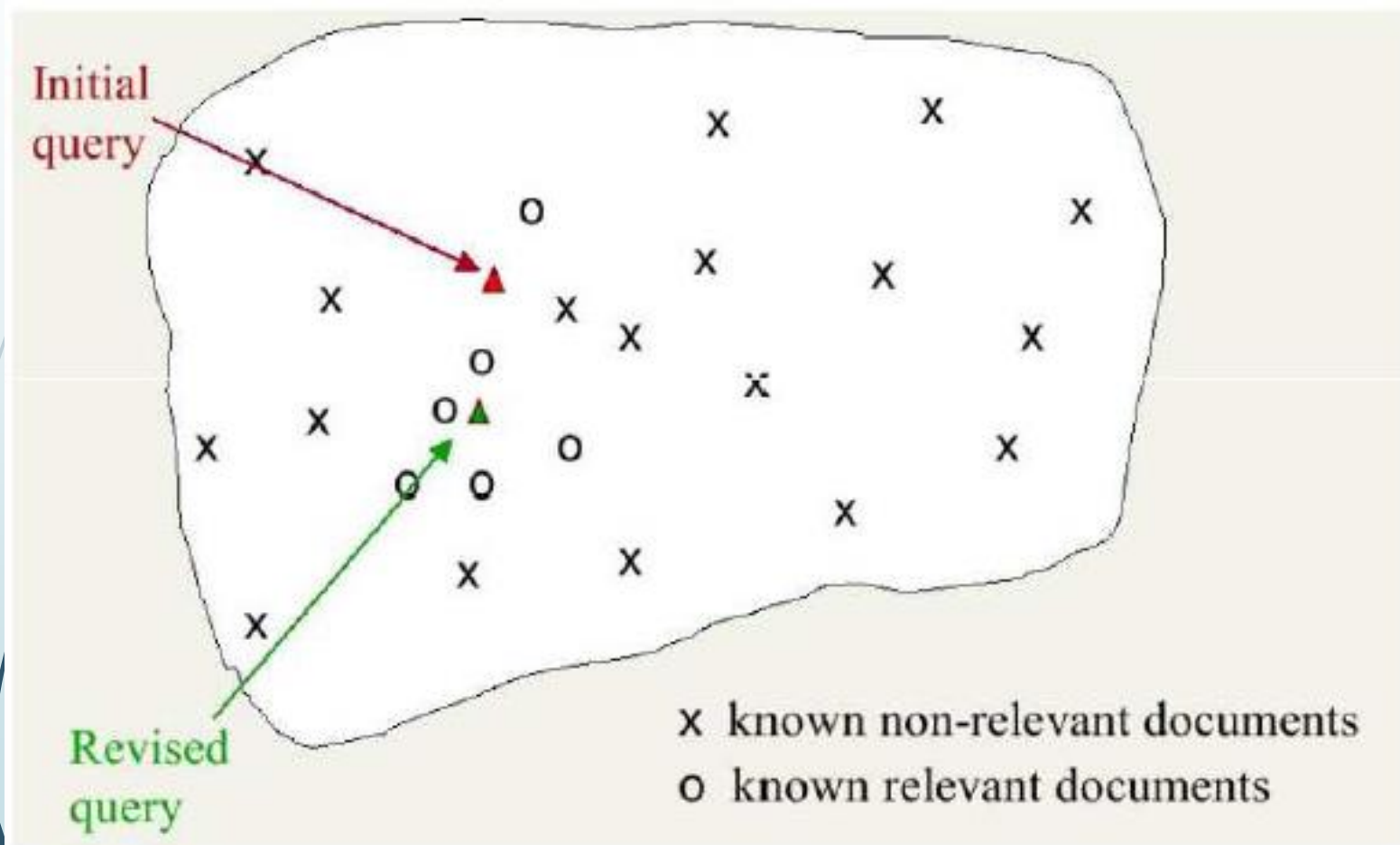
- Outra ideia: reformulação da consulta à partir de informações do usuário sobre a relevância de documentos.
- 1. Usuário submete a consulta original;
- 2. Uma lista de documentos recuperados é apresentada ao usuário;
- 3. O usuário examina os documentos e marca aqueles que são relevantes;
- 4. Com base na informação fornecida pelo usuário, o sistema computa uma nova consulta;
- 5. A nova consulta é submetida ao sistema.



# Realimentação de relevância explícita

- O objetivo principal consiste em:
  - Selecionar termos importantes dos documentos que foram identificados como relevantes pelos usuários;
  - Aumentar a importância desses termos em uma nova formulação da consulta.
- Espera-se que a nova consulta seja movida para mais perto dos documentos relevantes e para mais longe dos documentos não relevantes.

# Realimentação de relevância explícita

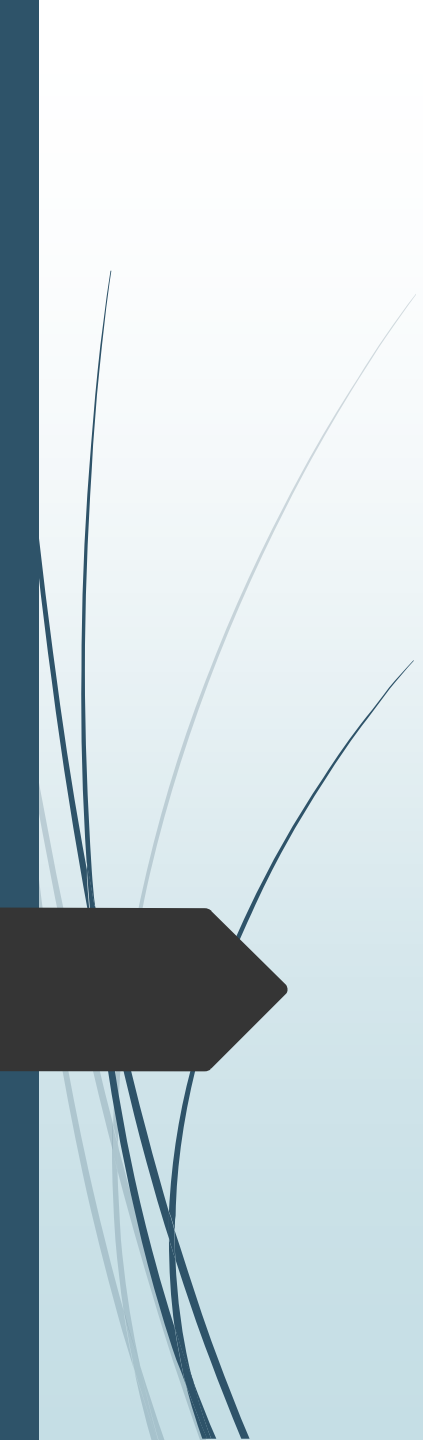




# Realimentação de relevância explícita

- Características:
- Evita que o usuário tenha que se envolver com o processo de reformulação da consulta (ele somente precisa fornecer julgamentos de relevância para os documentos);
- Divide a tarefa da busca em uma sequência de pequenos passos que são mais fáceis de aprender.





# Métodos de realimentação de relevância implícitos



# Realimentação implícita

## Análise global

- Documentos recuperados para uma determinada consulta  $q$  são examinados para determinar os termos para a expansão da consulta;
- Semelhante a um ciclo de realimentação de relevância, mas feito sem o envolvimento do usuário;
- A consulta é modificada baseada em alguma característica global da coleção, isto é, um recurso que não depende da consulta.



# Realimentação implícita

## Análise global

- Informação principal utilizada: sinônimos (ou quase sinônimos);
- Uma publicação ou base de dados que lista esse tipo de sinônimo é conhecido como **tesauro**;
- Existem dois tipos de tesauro:
  - Criados manualmente;
  - Criados automaticamente.

# Realimentação implícita

## Análise global

Google

instituto federal ms



instituto federal ms

instituto federal muzambinho

instituto federal montes claros

instituto federal minas gerais

Aproximadamente 2.040.000 resultados (0,39 segundos)

Google

universidade f



universidade federal de viçosa

universidade federal de juiz de fora

universidade federal de uberlândia

universidade federal de lavras

Pressione "Enter" para pesquisar



# Análise global

## Tipos de expansão de consultas

- Tesouro construído manualmente (mantido por editores, exemplo UNESCO);
- Tesouro construído automaticamente (baseado em estatísticas de coocorrências, por exemplo);
- Equivalência de consulta baseada em mineração do histórico de consultas (muito comum na Web como no exemplo do slide anterior).



# Expansão de consultas com tesouros

- Para cada termo  $t$  na consulta, a ideia consiste em expandi-la com palavras relacionadas ao termo  $t$  presentes no tesouro;
  - Ex. HOSPITAL – MÉDICO;
- Em geral aumenta a recuperação mas pode diminuir a precisão com a presença de termos ambíguos;
- Utilizado amplamente em motores de busca especializados para ciência e engenharia;
- É muito caro criar e manter um tesouro manualmente.



# Exemplo de tesauro manual UNESCO

- Tesauro utilizado para o controle e indexação de termos das áreas de Educação, Cultura, Ciências Naturais, Sociais e Humanas
- (<http://databases.unesco.org/thesaurus/>)



# UNESCO Thesaurus

Content language English ▾  ✕ Search

Alphabetical Hierarchy Groups

A B C D E F G H I J K L M N O  
P Q R S T U V W X Y Z

*International maritime law* → Law of the sea  
*International migration* → Migration  
*International monetary systems*  
*International organizations*  
*International politics*  
*International recommendations* → International instruments  
*International regional organizations* → Regional organizations  
*International relations*  
*International schools*  
*International security*  
*International solidarity*  
*International standards* → Standards  
*International students* → Foreign students  
*International studies* → International schools  
*International tensions*  
*International trade*  
*International training programmes*  
*International transport*  
*International travel* → Travel abroad  
*International understanding education* → International education  
*International universities*  
*International voluntary services*  
*Internationalism*  
*Internet*

## Vocabulary information

TITLE	UNESCO Thesaurus
DESCRIPTION	The UNESCO Thesaurus is a controlled and structured list of terms used in subject analysis and retrieval of documents and publications in the fields of education, culture, natural sciences, social and human sciences, communication and information. Continuously enriched and updated, its multidisciplinary terminology reflects the evolution of UNESCO's programmes and activities.
IDENTIFIER	<a href="http://vocabularies.unesco.org/thesaurus">http://vocabularies.unesco.org/thesaurus</a>
PUBLISHER	UNESCO
RIGHTS HOLDER	UNESCO
RIGHTS	CC-BY-SA
LICENSE	<a href="http://creativecommons.org/licenses/by-sa/3.0/igo/">http://creativecommons.org/licenses/by-sa/3.0/igo/</a>
CREATED	Saturday, January 1, 1977 00:00:00
LAST	Thursday, June 24, 2021 14:43:57





# Exemplo de tesauro manual INEP

- Vocabulário controlado que reúne termos e conceitos extraídos de documentos analisados no Centro de Informação e Biblioteca em Educação.
  - <http://inep.gov.br/thesaurus-brasileiro-da-educacao>

Cibec - Centro de Informação e Biblioteca  
em Educação

Pesquisa ao Acervo

Thesaurus Brasileiro da Educação

Rede de Especialistas

Política de Desenvolvimento de Coleções

Bibliografia Brasileira de Educação

[Página Inicial](#) > [Sobre o Inep](#) > Cibec - Centro de Informação e Biblioteca em Educação

# Thesaurus Brasileiro da Educação

O *thesaurus* é um instrumento que reúne termos escolhidos a partir de uma estrutura conceitual previamente estabelecida e destinados à indexação e à recuperação de documentos e informações num determinado campo do saber.

O [Thesaurus Brasileiro da Educação \(Brased\)](#) é um vocabulário controlado que reúne termos e conceitos, extraídos de documentos analisados no Centro de Informação e Biblioteca em Educação (Cibec), relacionados entre si a partir de uma estrutura conceitual da área. Estes termos, chamados descritores, são destinados à indexação e à recuperação de informações. Não é simplesmente um dicionário, mas um instrumento que garante aos documentalistas e pesquisadores o processamento e a busca destas informações.

Pesquisa *Thesaurus*


Acesso à pesquisa do *Thesaurus* Brasileiro da Educação






## Pesquisa Tesouros

Selecione outras pesquisas ▾



Pesquisar

Limpar

 Opções de consulta

Buscar por: Todos ▾

Registros por página: 20 ▾

### Resultados "7"

- Instituto de Educação (TG: Instituição Superior de Educação)
- Instituto de Pesquisa Científica e Tecnológica (TG: Instituição de Pesquisa)
- Instituto Politécnico (TG: Instituição de Formação Profissional Superior)
- Instituto Pós-Secundário (TG: Instituição Não-Universitária)
- Instituto Superior de Educação (TG: Instituição Superior de Educação)
- Instituto Tecnológico (TG: Instituição de Formação Profissional Superior)
- Instituto Universitário (TG: Unidades Universitárias)



# Geração automática de tesauro

- Tentar gerar um tesauro automaticamente ao analisar a distribuição das palavras em documentos;
  - Ideia básica: similaridade entre duas palavras;
- Definição 1: Duas palavras são similares se elas coocorrem com as mesmas palavras:
  - “carro” é similar a “motocicleta” pois ambos ocorrem com “estrada”, “gasolina” e “carteira”;
- Definição 2: duas palavras são similares se elas ocorrem em uma da relação gramatical com as mesmas palavras:
  - Maça é similar a pera pois em ambos os casos podemos colher, descascar, comer, preparar...




# Realimentação implícita

## Análise do contexto local

- Usa grupos de substantivos selecionados a partir dos documentos no topo do ranking;
- Em vez de documentos, são usadas passagens de tamanho fixo do documento (por exemplo, exemplo, 300 palavras) para a determinação das coocorrências entre termos;



# ATIVIDADES DA SEMANA

- 
- DESENVOLVIMENTO DO TRABALHO;
  - RESOLUÇÃO DA LISTA III;

# Bom Início de Semana!

*Que seja repleto de fé, boas notícias,  
sorrisos, produtividade e amor*

