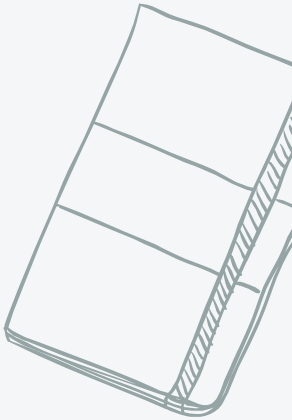
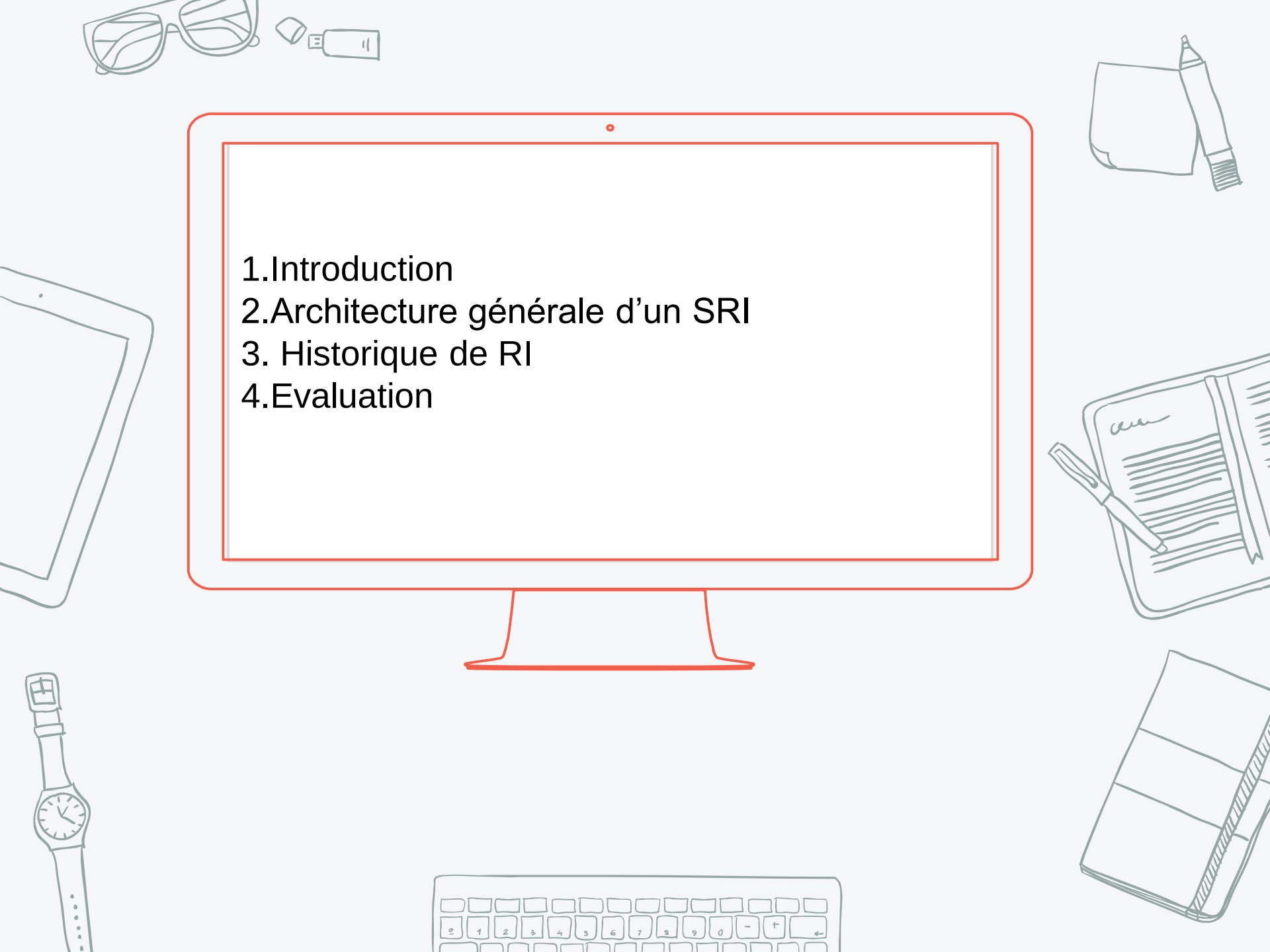
Indexation en recherche d’information pour le contenu



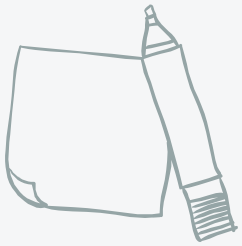


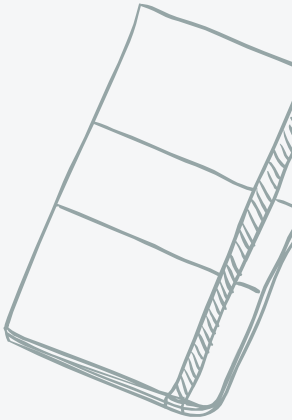

1.Introduction
2.Architecture générale d’un SRI
3. Historique de RI
4.Evaluation



Qu'est-ce que c'est un système de recherche d’information ?

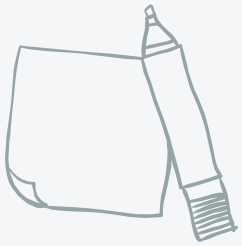





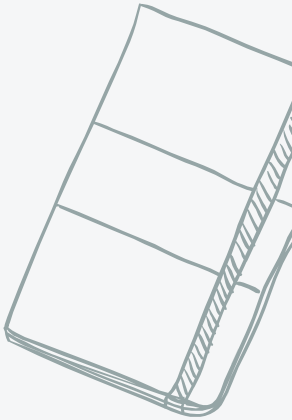



✕-Un système qui permet de retrouver une *information pertinente* par rapport à une *requête* dans une grande collection de *documents*.

✕-La recherche d’information (RI) ou (SRI) selon Salton :La branche de l’informatique qui consiste à acquérir, organiser, stocker, rechercher et sélectionner l’information.

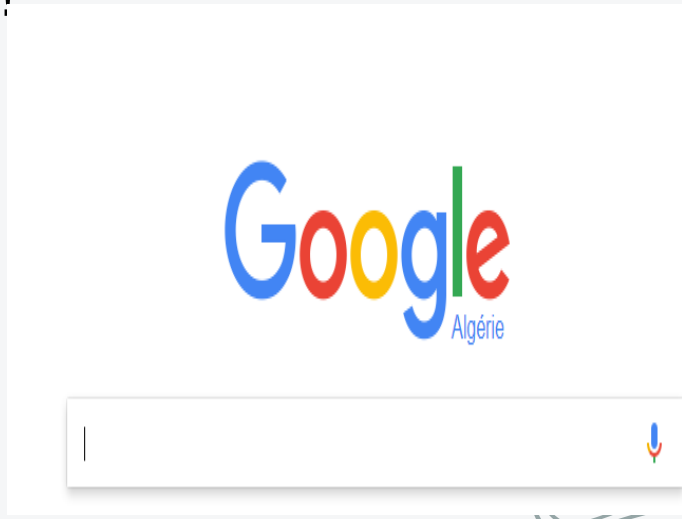




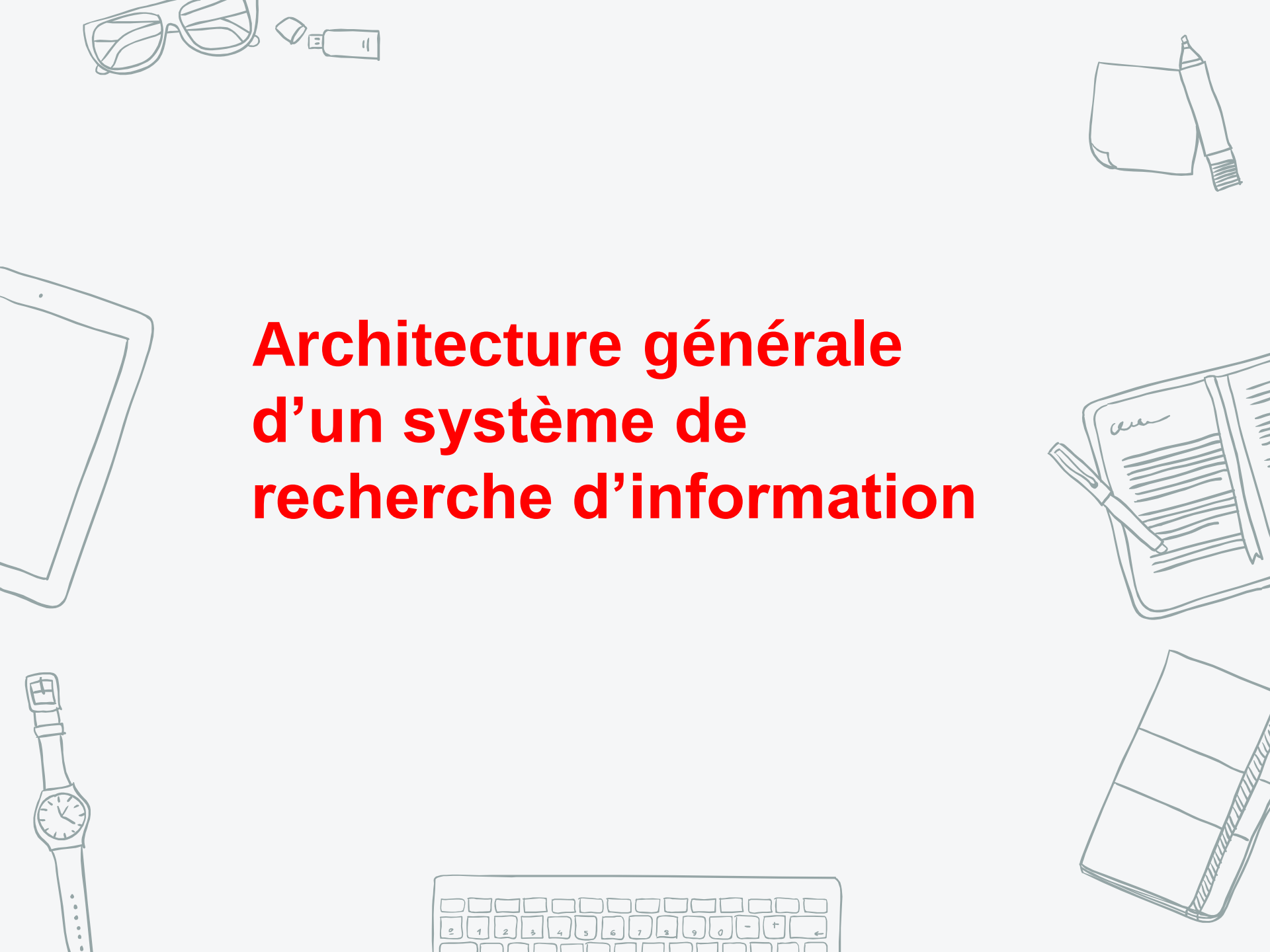

RI: Elle prend plusieurs terminologies

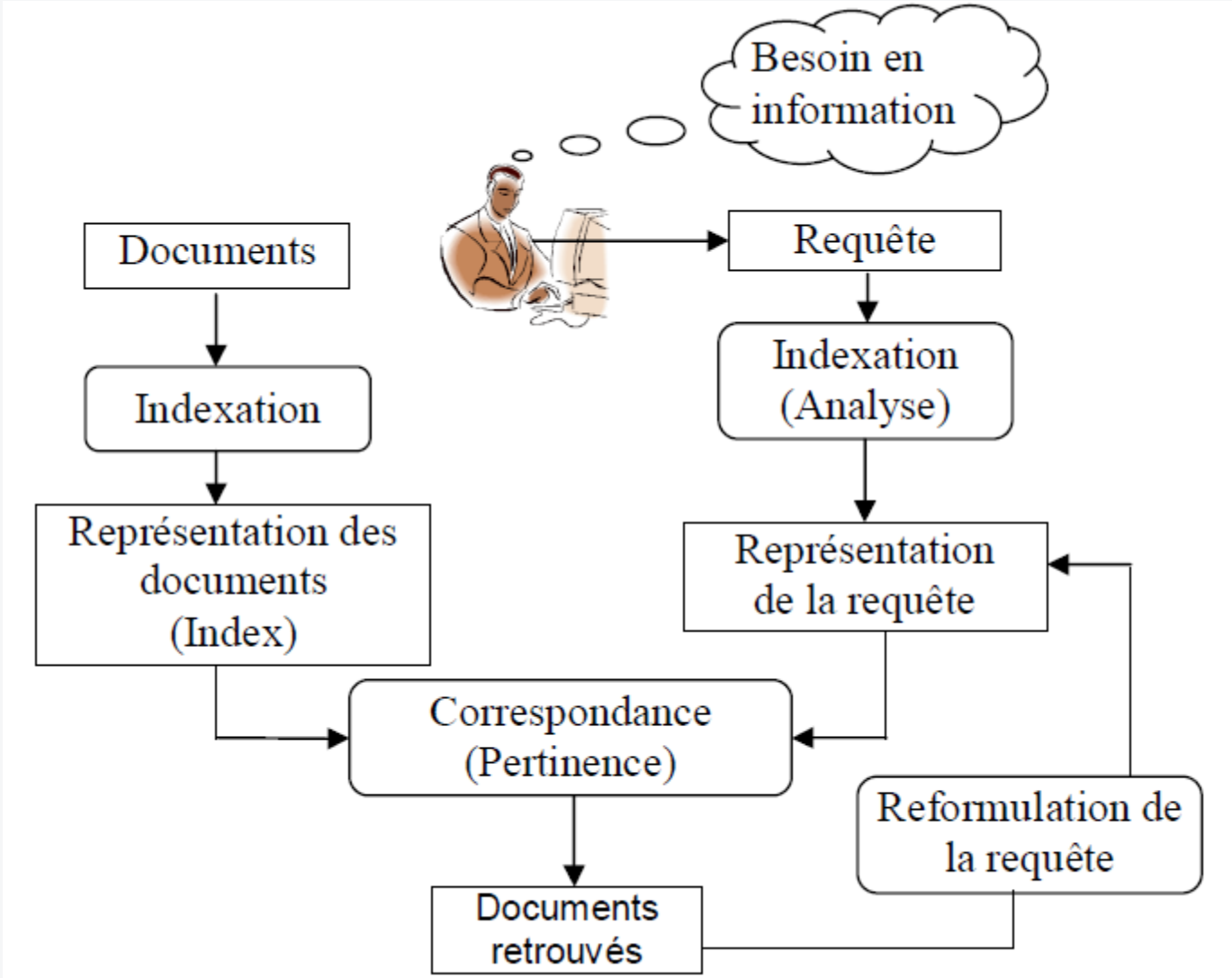
-Recherche d’information
-Informatique documentaire
-Information retrieval
-document retrieval

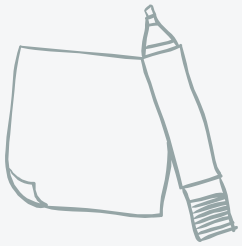
Domaines d’application!!!!!!



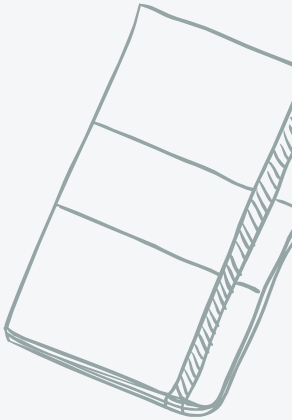
Architecture générale d’un système de recherche d’information




Besoin en information
Documents
Indexation
Représentation des documents (Index)
Requête
Indexation (Analyse)
Représentation de la requête
Correspondance (Pertinence)
Reformulation de la requête
Documents retrouvés






Approches possible pour réaliser une SRI

1. Une approche très naïve
2. Une approche basée sur une indexation

Historique de la RI

1940 : Avec la naissance des ordinateurs, la RI se concentrait sur les applications dans des bibliothèques. Depuis le début de ces études, la notion de pertinence a toujours été un objet.

• 1950 : Début de petites expérimentations en utilisant des petites collections de documents (références bibliographiques). Le modèle utilisé est le modèle booléen.

• 1960-1970 : Expérimentations plus larges ont été menées. On a développé une méthodologie d'évaluation du système qui est aussi utilisée maintenant dans d'autres domaines (des corpus de test ont été conçus pour évaluer des systèmes différents).

• 1970 : Développement du système SMART. Les travaux sur ce système a été dirigés par G. Salton. Certains nouvelles techniques ont été implantées et expérimentées pour la première fois dans ce système (par exemple, le modèle vectoriel et la technique de relevance feedback).

Du côté de modèle, il y a aussi beaucoup de développements sur le modèle probabiliste.

1980 : Les travaux sur la RI ont été influencés par l’avènement de l'intelligence artificielle.

Ainsi, on tentait d'intégrer des techniques de l'IA en RI, par exemple, système expert pour la RI, etc.

• 1990 : Internet à propulser la RI en avant scène de beaucoup d'applications.

La venue de l'Internet a aussi modifié la RI. La problématique est élargie. Par exemple, on traite maintenant plus souvent des documents multimédia qu'avant. Cependant, les techniques de base utilisées dans les moteurs de recherche sur le web restent identiques.

Années 2000

•Analyse de liens pour la recherche d'informations sur le web (google)

•Réponses à des questions (TREC QA track)

• indexation et recherche d'informations multimedia (image, video, audio et musique)

Recherche d'information multilingue (CLEF, NTCIR, DARPA, Tides)

Thanks!

Any questions?

You can find me at:
Sarah_Benmazou@live.fr








Introduction

•Avec l’explosion de la masse de données multimédia, le développement d’application et de moteurs de recherche pour l’exploiter devient crucial.

•Un document multimédia est toute unité qui peut être retrouvée par le système (image, vidéo…).

•Avec l’expansion de l’informatique et du multimédia, une problématique nouvelle est apparue: Gérer les quantités énormes et croissantes d’images numériques.

Une estimation

-180 millions images (indexable)soit environ 3 téraoctets, étaient réparties sur 800 millions de page disponibles sur internet.

-Le développement récent de la toile laisse imaginer les chiffres actuels.

-Une banque de données Corbis, compte 65 millions images dont 2 millions disponibles en ligne, ces images aux thèmes varies(sport, actualité, art,histoire,etc),parfois œuvres de photographes renommés, sont issues de diverses collections publiques ou privées du monde entier.

-L’image sous toutes ses formes (photographie, vidéo, graphique ou dessin numérique) est aujourd'hui a disposition des professionnels mais aussi du grand public.

Nous distinguons trois grande origines à ce phénomène:

1.La démocratisation de l’informatique, l’évolution rapide des performances et de la capacité de stockage des machines facilite le traitement et l’accè à l’information;

2.Les progrès réalisés dans le domaine des réseaux de télécommunication permettent la transition et le partage de l’information numérique aussi bien localement que mondialement;

3.Les images fixes et les séquences d’images sont en général compressées puis archivées dans les base de données, généralistes ou spécialisées, accessible par les réseaux de télécommunication.

Pour la réutilisation de ces données, il est donc capital de développer des outils de recherche. Ils sont généralement directement lies aux outils d’archivage, voire de compression des documents numériques.

Problèmes liés à l’indexation des images

La classification

Le catalogage

L’indexation

Le catalogage

- Consiste à décrire un document, quel que soit son format, permettant d'une part de l'identifier de façon unique et d'autre part de le repérer par le biais d'une caractéristique qui n'a pas rapport à son contenu (numéro ISBN, nom auteur…)

La classification

•Permet de placer un document, après avoir analysé le contenu de façon générale, dans l’ensemble des documents qui traitent le même sujet.

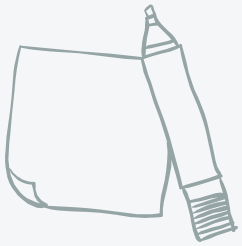
•Le document est ici considéré comme une entité. C’est un peu comme placer le document dans une boite étiquetée (animaux, Meubles….)

L’indexation

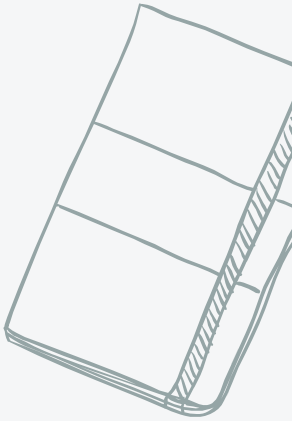
- On ne considère plus le document comme une entité distincte mais on considère plutôt les éléments d’information qui s’y trouvent.

- Le but de l’indexation est toujours de créer des regroupements de documents sur un même sujet, mais la description est plus précise.

Indexation trop spécialisée






•John Sunderland mentionne que l’indexeur de collections d’images, dans le domaine de l’analyse du sujet et de l’iconographie, fait face à des problèmes historiques, pratiques et théoriques.

Les collections d’images n’ont pas de système d’organisation universellement acceptés. Donc , à cause de leur organisation, la recherche dans de telles collections est presque exclusivement réservée à des personnes versées dans la lecture d’iconographie.

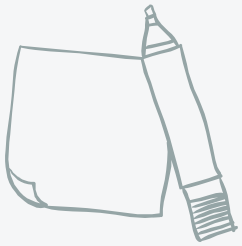
Mais de plus en plus, des gens de divers domaines consultent maintenant ces collections (sociologues, musiciens, publicistes, designers graphique.etc) pour trouver dont il ont besoin soit pour une annonce publicitaire ou une jaquette de livre.

Le problème majeur réside dans le fait que ces nouveaux usagers, qui n’ont pas nécessairement de connaissances dans le domaine de l’iconographie, ne peuvent facilement consulter ces collections avec les systèmes actuels de repérage des images en art. L’accès de ce genre de collections est donc limité…..

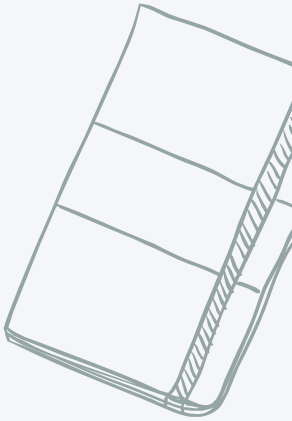
Q:La question demeure comment peut-on organiser la collection pour satisfaire tout le monde?

Polysémie de l’image

Selon Sara shatford Layne « The delight and frustration of pictorial ressources is that a picture can mean different things to different people »






Selon Sunderland, un des grands problèmes de l'image est qu'elle peut avoir une infinité de significations. Il a donc mené une petite enquête ou il à demandé à :

Un enfant de 12 ans : il a décrit les éléments se trouvant dans l'image: une femme à genou tenant un enfant. L'enfant à un trou dans sa main. Un homme tient la main de l'enfant et un clou ou quelque chose. Il y a des copeaux de bois sur le plancher, un jeune garçon avec un bol dans ses mains, etc.

Un profane: C'est une image religieuse ou est représentée la famille Sainte dans l'atelier de Joseph. IL mentionne aussi le fait qu'il est évident que la famille est heureuse et que l'on voit la relation d'amour et de bonheur qui existe entre le père, la mère et l'enfant.

Historien d'art: a identifié l'artiste, le titre le l'œuvre, sa date d'exécution, l'endroit où se trouvait le tableau, le style, le sujet(moment prophétique pour Jésus lorsqu'il se blesse avec un clou) et il continu son interprétation en expliquant la symbolique de l'œuvre….

Choix des termes

Selon Michael Krause: un des grands problèmes porte sur le choix des termes pour l'indexation des images.

La difficulté ne consiste pas à choisir le système d'indexation pour organiser les images mais plutôt de son application

Comment définir les sujets

les termes même qu'il faudrait choisir.

Il existe deux types d'indexation, le " hard indexing " ou le " ofness " d'une image (l'image est de …), i.e. ce que l'indexeur voit dans l'image, soit un chat ou une femme par exemple. Et le " soft indexing " ou le " aboutness " d'une image (l'image est à propos de …) qui porte sur la signification;

Mais Krause : les indexeurs ne veulent pas aborder la signification de l'image de peur de devenir subjectif au moment de l'indexation.

Ginette Bléry (1981) pense aussi qu'il faudrait aborder cette question de la subjectivité de l'image. Elle a effectué une expérience où elle a demandé à des personnes (elle ne mentionne pas le nombre de participants) d'indiquer à l'aide de mots abstraits leurs impressions pour chacune des images qu'on leur montre (ni la quantité ni le genre d'image n'étaient indiqués).

Des 438 réponses obtenues, elle a pu tirer quatorze couples d'opposition. Selon Bléry , il n'y a pas une infinité de possibilités de significations pour les images et il suffit d'identifier les émotions que peut susciter une image parce qu'en général elle est la même pour tout le monde.

Voici donc les quatorze couples d'opposition en question qui pourraient servir à décrire l'aspect subjectif d'une image :

Abstrait et Sensuel
Actif et Passif
Ancien et Moderne
Apaisant et Stimulant
Artificiel et Naturel
Beau et Laid
Sérieux et Frivole
Chaud et Frais
Coloré et Terne
Gai et Triste
Comique et Tragique
Décontracté et Angoissé
Érotique et Froid
Ordonné et Discordant

Les besoins des usagers

•Kevin Roddy : même si les images sont emmagasinées de façon intelligente, elles restent inaccessibles étant donné qu'elles possèdent trop peu de descripteurs.

•Le problème des descripteurs est qu'ils sont ambigus, arbitraires et portent à confusion ou à des désaccords.

•Il indique qu'il serait important que le système indique des valeurs aux items repérés par degré de rapprochement avec la requête.

Transfert de la signification

•Un autre grave problème dans le domaine de l'indexation des images est le transfert de la signification d'un médium visuel vers du verbal.

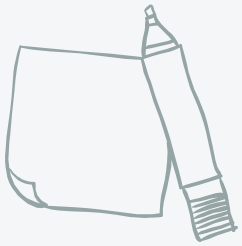
•On croit en effet que ce transfert cause la perte d'une partie de l'information. Surtout lorsque l'œuvre d'art ne contient pas d'information comme titre, description….

•En effet, Sunderland explique qu'il y a des images où l'on ne sait pas de quoi elle s'agit…..

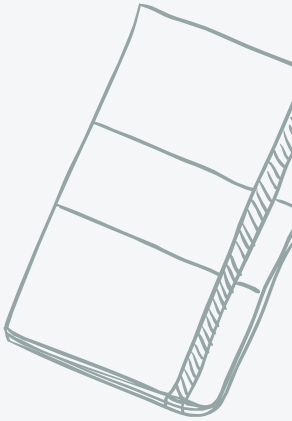
La figure suivante représente une famille partant en pique-nique un dimanche

En fait, elle représente des fermiers américains de l'Arkansas qui à cause de la fameuse crise des années 30 se voient obligés d'abandonner leur terre et de suivre l'exode vers la Californie.

Requêtes






Dans le passé, la problématique de la recherche d’images se résumait en une problématique de recherche de mots en se basant sur les attributs textuels des images tels que le nom de fichier.

Mes cette approche nécessite une entrée manuelle des mots définissant l’image (légende) et ne peut donc plus être appliqué aux flux, toujours croissant, d’arrivée des nouvelles images.

Il existe plusieurs types de requêtes:

-Parcours au hasard: la base est parcourue aléatoirement jusqu’à ce que l’utilisateur trouve l’image qui l’intéresse.

-Navigation par catégorie: les images sont classées par catégories. L’utilisateur donc directement choisir la catégorie dans laquelle il pense pouvoir trouver l’image.

-Recherche par mots clés: L’utilisateur entre des mots sensés représenter l’image recherché. Il dispose souvent d’une série de termes prédéfinis pour formuler sa requête.

-Recherche par l’exemple: Lors d’une telle requête, l’utilisateur fournit une image exemple et le logiciel recherche dans la base les images qui ressemblent à l’image exemple.

L’image exemple peut être (une photo de l’objet désiré ou une représentation crée par l’utilisateur lui-même-dessin ou image de synthèse par exemple).

Les deux premiers méthodes ci-dessus ne peuvent pas,de par leur fonctionnement, être utilisés pour une indexation et récupération automatique des images.

La recherche par mots clés, très présente dans l'indexation de documents textuels, ne semble pas vraiment adapté aux images.

Premièrement, il peut s'avérer très difficile de décrire une image en quelques mots, surtout si l'utilisateur n'a qu'une vague idée de ce qu'il désir.

Deuxièment, une requête par mots clés nécessite une longue et couteuse phase manuelle d'indexation des images de la base avec des mots clés les définissant.

Cette indexation manuelle des images et devenue impraticable avec la taille toujours grandissante des bases de données multimédia actuelles.

La recherche par l'exemple est une nécessité majeur pour les utilisateurs vu qu'elle supprime le besoin d'exprimer la requête à l'aide de mot.

De plus, elle élimine l'exigence de la couteuse étape préliminaire d'indexation alphanumérique manuelle.

Avec cette requête, le problème d'indexation peut se transformer en une définition d'une distance entre images.

Indexation par le contenu

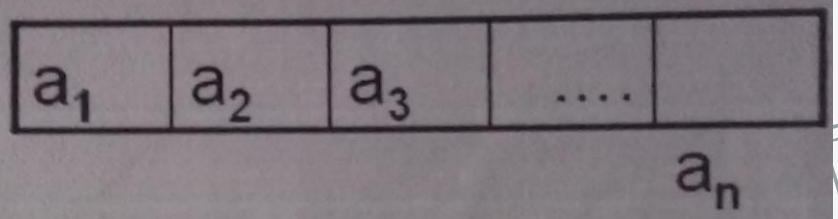
Plusieurs systèmes d’indexation et de recherche de documents multimédia par le contenu ont vu le jour. La majorité d’entre eux concernent les images.

Un système typique de recherche d’image par le contenu permet aux utilisateurs de formuler de requêtes en présentant un exemple du type de l’image recherchée, bien que certain offre d’autre solutions telles que la sélection dans une liste d’exemple.

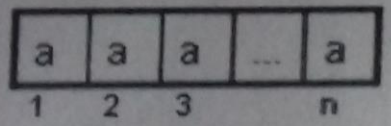
Le système identifie alors parmi la collection d’images celles qui correspondent le plus à l’image requête, et les affiche.

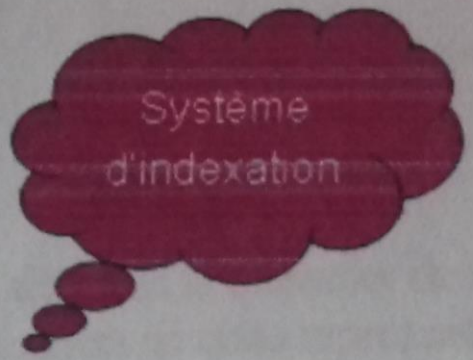
Le processus d’indexation par le contenu se résume en deux étapes:

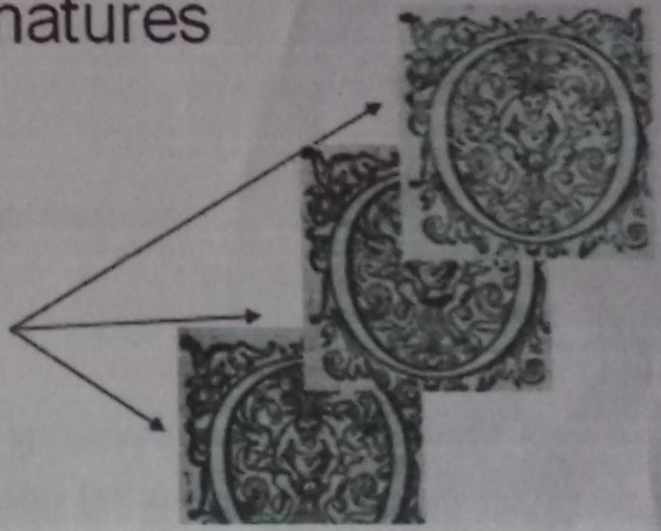
- Extraction de signatures



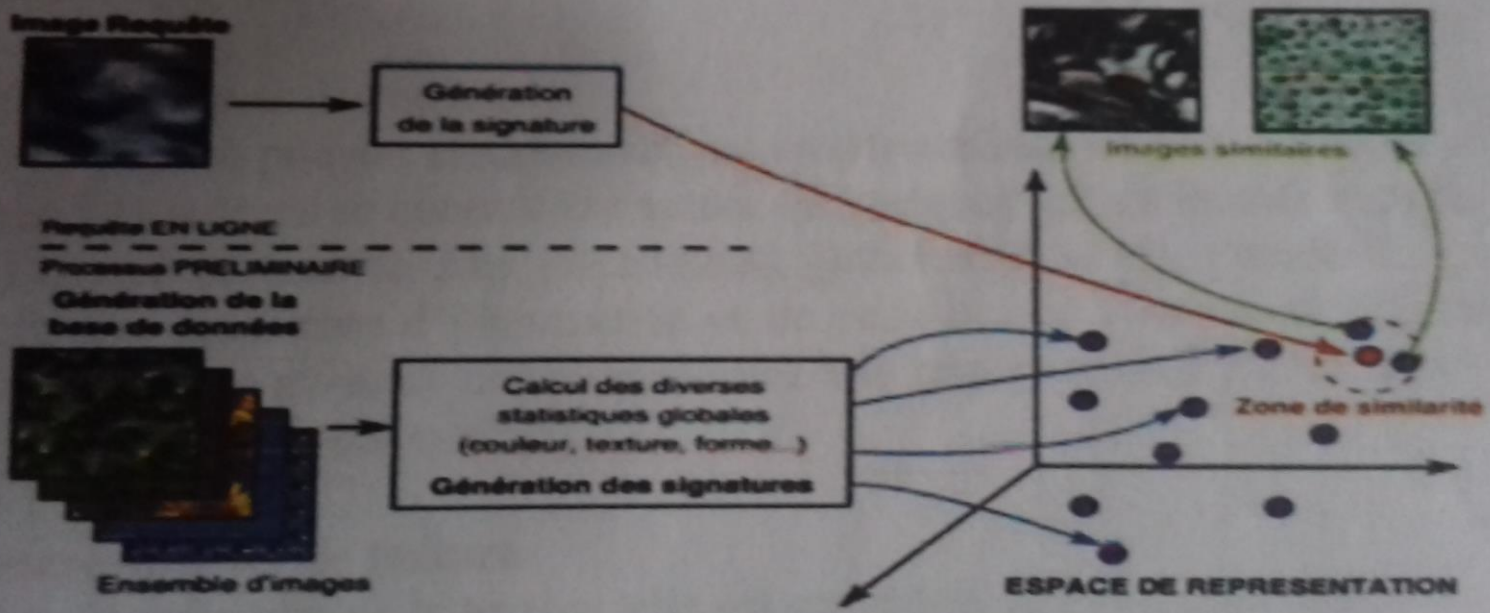
- Indexation à partir de ces signatures







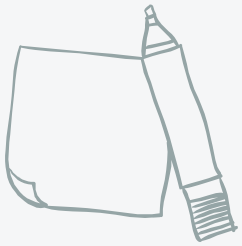
la démarche d'une recherche par l'exemple :



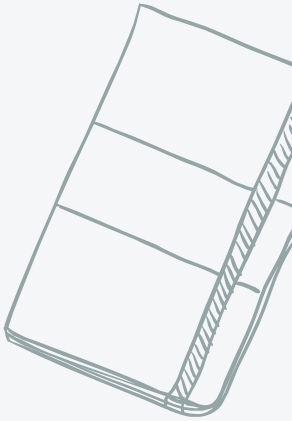
– Schéma d'indexation classique

On distingue deux types d’approches de recherche d’information multimédia par le contenu.

Bas niveau
Haut niveau






Dans le cadre de la recherche d’information multimédia, on assigne aux documents des concepts, ou termes sémantique.

Ce processus appelé « indexation sémantique » peut être réalisée de trois manières différentes:

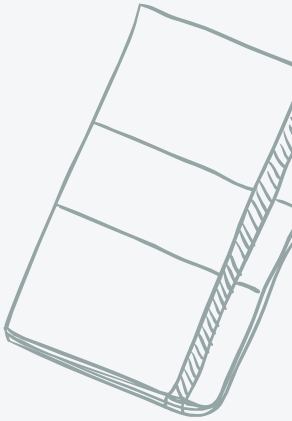
Manuelle

Automatique

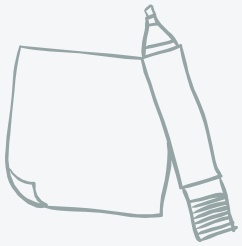
Semi-Automatique



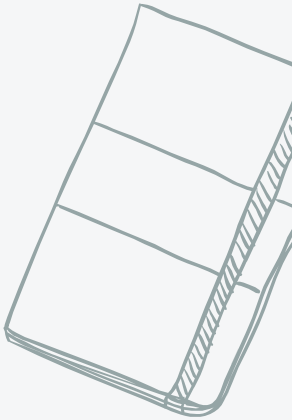



Difficultés et défis de l’indexation sémantique






Devant faire correspondre une représentation brute (de bas niveau) à une description conceptuelle ou sémantique, l'indexation automatique se heurte à un problème appelé "le fossé sémantique" (*semantic gap en anglais).*

Selon Ayache "*Le fossé sémantique est ce qui sépare les représentations brutes (tableaux de nombres) et sémantiques (concepts et relations) d'un document numérique".*

Smeulders et al… donnent une autre définition : "*Le fossé sémantique est le manque de concordance entre les informations que la machine peut extraire d'un document numérique et des interprétations humaines".*

Les mêmes auteurs évoquent un autre type de fossé dit "fossé sensoriel", qui est défini comme le fossé existant entre le monde réel 3D et sa représentation en une image 2D.

Lors de l’acquisition des images et vidéos, cette projection vers un espace 2D provoque une grande perte d’informations.

Cela mène à une représentation de bas niveau, pas très efficace.

-Les caractéristiques visuelles de bas niveau ne parviennent souvent pas à décrire les concepts sémantiques de haut niveau.

D’autre part, la prise en photo ou en vidéo d’une même scène ou même objet par des dispositifs différents, dans des situations différentes, mène à des documents multimédia différents en termes de :

- Luminosité/illumination (a) : La présence d’ombre ou de halos sur un objet provoque parfois des occultations partielles ;

(a) Changement de luminosité

– Taille de l’objet ou l’´echelle de l’image : (b)

(b) Changement d’´echelle (taille de l’objet), exemple de vélo

– L’angle de prise de vue (c): un même objet pris de différents angles de vues peut apparaitre sous des formes variées, parfois n’aidant pas à savoir qu’il s’agit du même objet ;

(c) Changement de l’angle de prise de vue

Le fossé sémantique est également accentué à cause de :
– La variabilité visuelle intra-classe(d) : les objets d'une même classe n'ont pas toujours les même caractéristiques visuelles, on parle de "variations visuelles" ou de "multiples représentations d'un même objet" ;

(d) Différence entre les caractéristiques visuelles d'un même objet (exemple de la classe *avion)*

– La variabilité visuelle inter-classes (figure(e) : des descriptions visuelles similaires peuvent concerner deux concepts qui n'ont rien à voir l'un à l'autre (des descripteur visuels similaires concernant deux objets différents) ;
Tous ces problèmes compliquent la tache à la machine pour déduire que ces contenus numériques qu'elle manipule correspondent au concept recherché ou non.
Franchir le fossé sémantique constitue une des difficultés majeures d'un système automatique d'annotation/indexation de documents multimédia. Les communautés de la vision par ordinateur et de l'indexation automatique continuent de traiter ce problème. Beaucoup de travaux ont été réalisés dans le but d'augmenter la corrélation entre des contenus visuels similaires sémantiquement en proposant de bons descripteurs. D'autres méthodes d'apprentissage automatique ont été proposées, qui permettent de faire correspondre plus efficacement les caractéristiques de bas niveau à des descriptions sémantiques ou des concepts.

(e) Ambiguïté visuelle. Un même contenu visuel ou deux contenus visuels similaires peut référer à deux sens différents

Merci