Intérêts d'un Corpus issu du Web pour les Systèmes Question-Réponse

Véronique Moriceau, Farida Aouladomar

Plan

- Cadre et problématique
- Constitution et exploitation d'un corpus Web
- Illustration pour les questions-réponses :
  - factoïdes (entités)
  - procédurales
- Conclusion

Cadre et problématique

- Aujourd'hui : très grand nombre de données sur le Web
- Rechercher une réponse à une question :
  - dictionnaires /encyclopédies : réponse unique, synthétique, cohérente
  - systèmes de recherche d'informations / moteurs de recherche : ensemble de liens vers des pages Web et/ou des extraits de ces pages traitant du thème de la requête

⇒ l'utilisateur doit sélectionner les pages les plus intéressantes et rechercher au sein des textes la réponse à sa question.

⇒ démarche laborieuse et longue !

Cadre et problématique

(cf. O Jenhani)

Trop de réponses !

Pas toutes pertinentes !

activer sa freebox - Recherche Google - Microsoft Internet Explorer

Fichier Edition Affichage Favoris Outils ?

Précédente Rechercher Favoris Média

Adresse http://www.google.fr/search?hl=fr&q=activer+sa+freebox&meta=

Liens Community Hotmail Internet Instantanée Mon Presario

Google

Web Images Groupes Annuaire Actualités plus

activer sa freebox

Rechercher dans : Web Pages francophones

Web

Activer sa freebox pour pouvoire utiliser freeplayer/payant ...
Discussions au sujet de **Activer sa freebox** pour pouvoire utiliser freeplayer/payant? dans le forum Internet & Réseaux, **activer freebox** pouvoire utiliser ...
www.infos-du-net.com/forum/ 172606-8-**activer-freebox**-pouvoire-utiliser-freeplayer-payant - 25k - En cache - Pages similaires

freebox en mode routeur : configuration pas à pas (eMule, FTP, P2P)
Voici comment **activer** le mode routeur de la **freebox** mais également comment bien le ... le serveur pour lui soumettre des demandes, et attendent **sa** réponse. ...
www.memoclic.com/formation_1574/commentconfigurerla**freebox**enmoderouteur.html - 23k - En cache - Pages similaires

Forum Freebox - SONDAGE : qui est heureux de sa FREEBOX ?
La question pour Free n'est pas "qui est heureux de **sa FREEBOX** ?" Le PB c'est pas le fonctionnement de la **Freebox**, c'est de la recevoir et de faire **activer** ...
www.dslvalley.com/forum/forum_ read.php3?fqID=28684&ID=0303 - 48k - En cache - Pages similaires

Freebox - Wikipédia
La **Freebox** est prêtée aux abonnés, **sa** valeur étant de 400 € d'après l'opérateur. ... La télécommande fournie avec la **freebox** permet non seulement d'**activer** ...
fr.wikipedia.org/wiki/Freebox - 46k - En cache - Pages similaires

Boxbi, l'amie des Freebox, AOLBox, NeufBox, LiveBox et Cie ...
Pour la **Freebox** il faut enclencher les deux sélecteurs. ... tâtonnement qui n'est franchement pas facile surtout s'il faut "tester" pour **activer sa** Boxbi ! ...
www.informanews.net/scripts/article.php/170.htm - 29k - 5 mar 2006 - En cache - Pages similaires

Activer sa freebox pour pouvoire utiliser freeplayer/payant? - Internet & Réseaux - Microsoft Internet Explorer

Fichier Edition Affichage Favoris Outils ?

Précédente Rechercher Favoris Média

Adresse http://www.infos-du-net.com/forum/172606-8-activer-freebox-pouvoire-utiliser

Liens Community Hotmail Internet Instantanée Mon Presario Music Central

ox pour pouvoire utiliser freeplayer/payant?

de Cegetel

Conversation

Activer sa freebox pour pouvoire utiliser freeplayer/payant? alerte

voila je voudrai savoire si c payant d'activer sa freebox pour utiliser freeplayer parce que quand je ve lactiver il il dise si jai pri consience des tarif et tout donc jai pas été jusqua létape 2.

http://adsl.free.fr/tv/freeplayer/cgv_freeplayer.html

profil jabber

Terminé Internet

Internet

Cadre et problématique

- Amélioration : les systèmes question-réponse (coopératifs)
- Recherchent un ensemble de pages traitant de la question posée (moteur d'extraction)
- Proposent à l'utilisateur une réponse unique, celle que le système juge la "meilleure".

Cadre et problématique

- Système question-réponse :
  - proposer à l'utilisateur une réponse correcte à partir d'un ensemble de pages Web traitant du thème de la requête.

- Il est nécessaire d'étudier en détail la forme et le contenu des textes du Web pour :
  - identifier les types de données à manipuler,
  - comprendre et résoudre les problèmes auxquels un système est confronté

Constitution d'un corpus Web

- Choix des questions :
  - Typologie classique des questions [Lehnert, 1978] :
    - questions atomiques (qui attendent des réponses de type entité),
    - questions narratives (qui attendent des réponses de type texte).
  - Méthode centrée sur les besoins des utilisateurs :
    - FAQ,
    - inventaires des questions les plus fréquemment posées sur le Web (générateurs de mots-clés : Google, Overture).
  - Enrichissement manuel :
    - questions des campagnes TREC ,
    - questions imaginées pour certains domaines sous-représentés.

Constitution d'un corpus Web

- Choix des réponses :
  - Les questions sont soumises à Google sous forme de mots-clés ou à QRISTAL sous forme de question en langue naturelle $\Rightarrow$ *ensemble de pages Web*
  - Nous n'avons considéré que les 20 premiers liens donnés par Google et QRISTAL, les liens suivants n'apportant pas de réponses pertinentes supplémentaires.
  - Démarche cohérente avec les habitudes des internautes : 80% des internautes ne consultent pas plus d'une dizaine de liens pour une requête.

Constitution d'un corpus Web

- Choix des réponses :
  - 1ère étape (manuelle) : ne garder parmi cet ensemble que les pages pertinentes qui proposent effectivement une réponse, même fausse, à la question :
    - pages qui contiennent le focus de la question et,
    - pages qui contiennent une information correspondant au type sémantique de la réponse attendue pour les questions atomiques,
    - pages contenant des textes dits procéduraux pour les questions procédurales.

Exploitation d'un corpus Web

- 1ère observations :
  - Problème de la quantité et la qualité des réponses obtenues pour le choix de la réponse finale.
  - Pour une même question, les réponses potentielles peuvent être au mieux redondantes, sémantiquement équivalentes mais aussi incohérentes, approximatives, etc.

⇒ L'analyse du corpus nous permet donc d'identifier, de quantifier et de typer ces problèmes pour pouvoir proposer des solutions au système.

Exploitation d'un corpus Web

- Choix d'une réponse parmi un ensemble de pages :
  - Privilégier la 1ère réponse donnée par le moteur de recherche n'est pas une solution acceptable :

    Google propose la page qui donne une réponse correcte en moyenne au 4ème ou 5ème lien (varie en fonction du type de question).

⇒ définir des méthodes qui élaborent une réponse appropriée en prenant en compte des critères mis en évidence lors de l'étude de corpus : les relations que les réponses ont entre elles, leur cohérence, leur structure, leur degré de précision, etc.

Illustration : les questions factoïdes

- Objectifs
  - Construire un corpus issu du Web pour analyser les différentes réponses proposées pour une même question
  - Les réponses sont-elles :
    - différentes ?
    - cohérentes ?
    - Précises / imprécises ?
    - …
  - À partir des différentes réponses obtenues sur le Web, proposer des méthodes pour élaborer une réponse correcte

Illustration : les questions factoïdes

- Constitution du corpus
  - Choix des questions :
    - questions qui n'attendent qu'une seule réponse

      Quand est mort Beethoven ? le 26 mars 1827
    - questions qui acceptent plusieurs réponses

      Où se trouve le parc Disneyland ? à Paris, Los Angeles, Tokyo, …

Type des réponses	**localisation**	**personne**	**numérique**	**temps**	**objet**	**TOTAL**
Nombre de questions	32	34	47	40	27	180

  - Choix des réponses : que les pages pertinentes

    Question : *Quand est mort Beethoven ?*

    Réponse : *Ludwig von **Beethoven** est bien **mort** des suites d'un empoisonnement au plomb*

    ⇒ pas conservée car ne donne pas d'information temporelle

Illustration : les questions factoïdes

- Exploitation du corpus
  - Identifier les principales relations existant entre un ensemble de réponses potentielles.
  - 4 relations définies dans [Webber et al, 2002] : l'équivalence, l'inclusion, l'agrégation et l'alternative + complémentarité.

Type des réponses	Nombre de questions	Équivalence	Inclusion	Agrégation	Alternative	Complémen-tarité
Localisation	32	11	11	13	3	5
Personne	34	13	0	5	11	0
Numérique	47	11	0	19	32	16
Temps	40	3	0	5	13	5
Objet	27	8	3	8	5	3
TOTAL	180	46	14	50	64	29

Illustration : les questions factoïdes

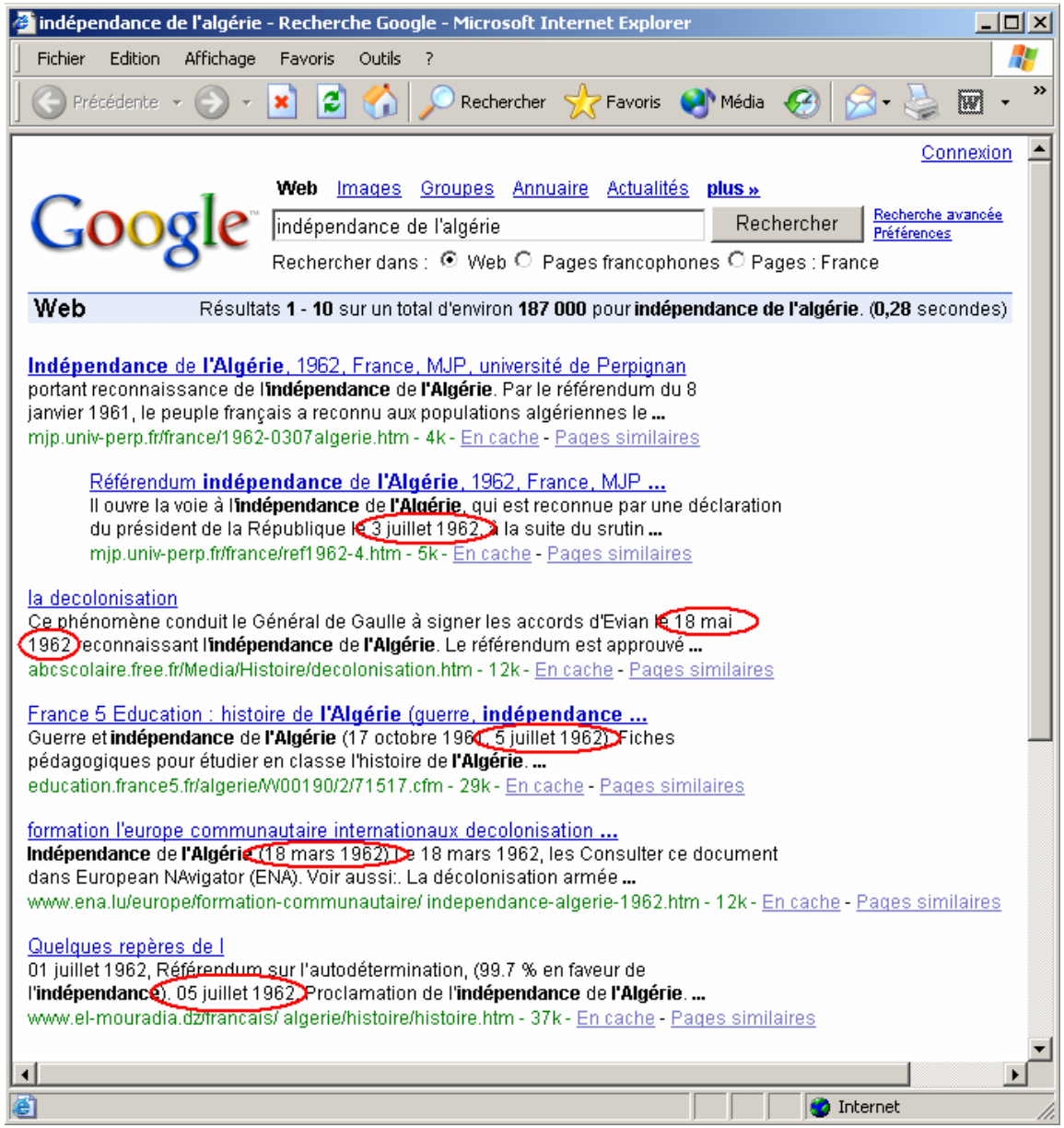


Illustration : les questions factoïdes

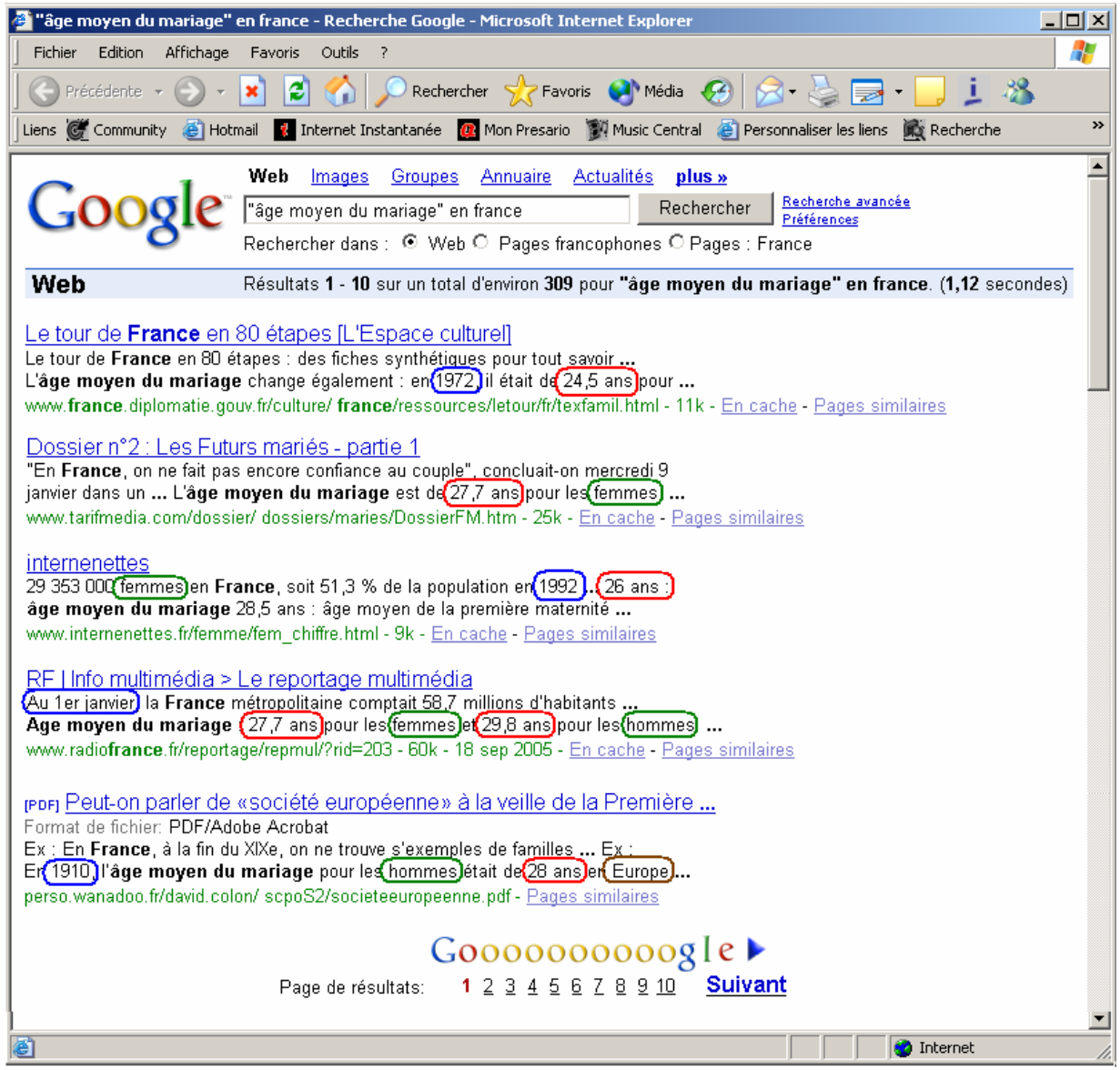


Illustration : les questions factoïdes

- Validation :
  - constituer un autre corpus de questions à soumettre au système pour l'évaluer

    ⇒ vérifier qu'il propose bien une réponse correcte

  - pour l'évaluation sur des questions temporelles :
    - 72 questions issues de la campagne TREC et des inventaires des requêtes temporelles les plus posées sur le Web,
    - questions portent sur différents types d'événements.

Réponse attendue	Evénement unique	Evénement itératif	TOTAL
Point (ponctuel)	18	18	36
Intervalle (duratif)	19	17	36
TOTAL	37	35	72

Illustration : les questions procédurales

- Constitution du corpus : Objectifs
  - Quelles formes pour les questions procédurales ?
    - Études des FAQ, liste de questions TREC, inventaires de requêtes sur le Web
  - Quelles formes pour les textes procéduraux ?
    - Comment repérer un texte procédural d'un texte non-procédural ?
    - Identifier les éléments informationnels des textes procéduraux utiles pour le mécanisme de recherche et génération des réponses.

Illustration : les questions procédurales

- Exploitation du corpus : les **Questions**
  - Questions procédurales = questions introduites par « comment »
    - *Comment vas-tu ?*
    - *Comment dit-on maison en espagnol ?*
    - *Comment est mort John?*
    - *Comment on mange le couscous au Maroc ?*
    - *Comment payer mes frais d'inscription à la fac ?*
    - *Comment changer une roue ?*

Illustration : les questions procédurales

– Autres formes de questions procédurales :

- Construction support : « que faire … », «quel + être + proposition»

que faire pour obtenir un visa ? quelles sont les démarches à effectuer pour obtenir un visa ?

- Le « comment » elliptique : mots clés

changer une roue, ajouter de la mémoire, assemblage ordinateur

- Inférence lexicale : mots clés

ordinateur en panne

– Questions attendant une réponse de type procédural :

- « est-il possible de …. »

est-il possible de créer un répertoire en Php ?

Illustration : les questions procédurales

- Constitution du corpus : textes procéduraux

Types de textes	Nb de textes procéduraux (inventaires)	Nb de textes procéduraux ajoutés	Nb total de textes dans le corpus
Communication / conseils	48	0	48
Domaines techniques	30	20	50
Santé	3	6	9
Recettes	0	10	10
Injonctions	0	7	7
Total	81	43	124

Illustration : les questions procédurales

- Exploitation du corpus : « procéduralité » d'un texte

  - Métrique « CATEG » : critères de surface
    - **Morphologie des verbes** (MSM) : impératifs, infinitifs
    - **Catégorie sémantique des verbes** (AV) : verbes d'actions
    - **Marques typo-dispositionnelles** (TF) : énumérations
    - **Connecteurs** (AM) : temporels, causalité

Illustration : les questions procédurales

- Exploitation du corpus : « procéduralité » d'un texte
  (Cédric Mayer)

Recettes	AM	MSM	TF	AV	CATEG
Textes procéduraux	1.21%	7.98%	5.49%	7.41%	**5.58**
Textes non procéduraux	0.70 %	3.21 %	1 %	5.02%	**2.42**

Domaines techniques	AM	MSM	TF	AV	CATEG
Textes procéduraux	1.80%	11.17%	4.36%	10.94%	**5.43**
Textes non procéduraux	1.20%	3.16%	1.66%	7.02%	**2.57**

Illustration : les questions procédurales

- Exploitation du corpus : structure des textes
  - spécificité des textes sur le web : textes pdf , html
  - hyperliens (références externes et internes)
  - isoler les informations pertinentes pour la réponse

Texte → title, (summary), (warning)+, (pre-requisites)+, (picture) + < objective.

Chercher sur CCM

OK ?

Assemblage

- Précautions à prendre!
- Préparer le boîtier
- Installer la carte-mère
- Mettre le processeur
- Enficher la RAM
- Insérer les cartes d'extension
- Brancher les disques
- Monter les périphériques dans les baies
- Configurer le BIOS

Livre sélectionné

(en association avec Amazon.fr)

4 janvier 200

summary

Objective

De l'assemblage d'un ordinateur (PC)

title

warning

- Toutes les opérations suivantes sont simples, mais une **ERREUR** (inversion de certains connecteurs) peut entraîner des **DEGATS IRREVERSIBLES** sur le matériel, ce guide est fait pour vous aider il n'est en rien une référence, c'est pourquoi CCM ne saurait être responsable de sa mauvaise utilisation.
- Pour toutes les opérations suivantes, il faut s'assurer d'avoir débranché le cordon d'alimentation du PC (puis, pour les puristes, de toucher le boîtier métallique d'une main, le sol de l'autre pour décharger l'électricité statique)!

Préparer le boîtier

Lorsque l'on se prépare à installer un PC, il faut vérifier que l'on possède un tournevis cruciforme, les vis dont on aura besoin, ainsi que les différents câbles et connecteurs.

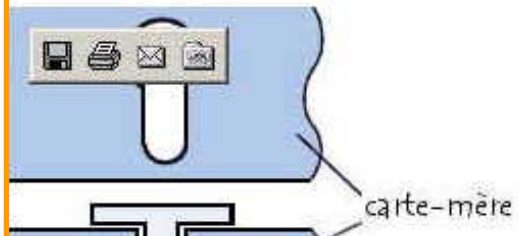
Il existe tout de même quelques règles simples à respecter :

- Ne jamais visser à fond !
- Ne jamais forcer !

La première étape consiste à ouvrir entièrement le boîtier, puis de le placer à plat sur une surface large où vous aurez suffisamment de place pour travailler confortablement, et enfin de retirer tous les caches en plastique des baies à l'avant du PC.

La carte-mère

Installation: La carte-mère se "clipse" dans le boîtier, des petits ergots sont en général fournis pour tenir la carte, puis on la visse pour la fixer.



Objective → {goal}< (warning)+, (picture)+, (pre-requisites), instruction sequences+ / objective

Préparer le boîtier

goal

Pre-requisites

Lorsque l'on se prépare à installer un PC, il faut vérifier que l'on possède un tournevis cruciforme, les vis dont on aura besoin, ainsi que les différents câbles et connecteurs.

Il existe tout de même quelques règles simples à respecter :

- Ne jamais visser à fond !
- Ne jamais forcer !

warning

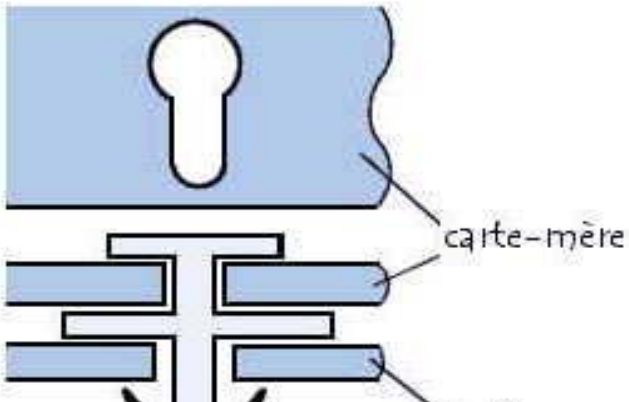
La première étape consiste à ouvrir entièrement le boîtier, puis de le placer à plat sur une surface large où vous aurez suffisamment de place pour travailler confortablement, et enfin de retirer tous les caches en plastique des baies à l'avant du PC.

NB: Un boîtier est généralement fourni avec un câble d'alimentation et un ensemble de vis. Les vis avec un petit pas de filetage s'installent généralement sur les lecteurs de disquettes et CD, alors que les vis avec un gros pas de filetage permettent de fixer les disques durs.

Instruction sequences

La carte-mère

Installation: La carte-mère se "clipse" dans le boîtier, des petits ergots sont en général fournis pour tenir la carte, puis on la visse pour la fixer.



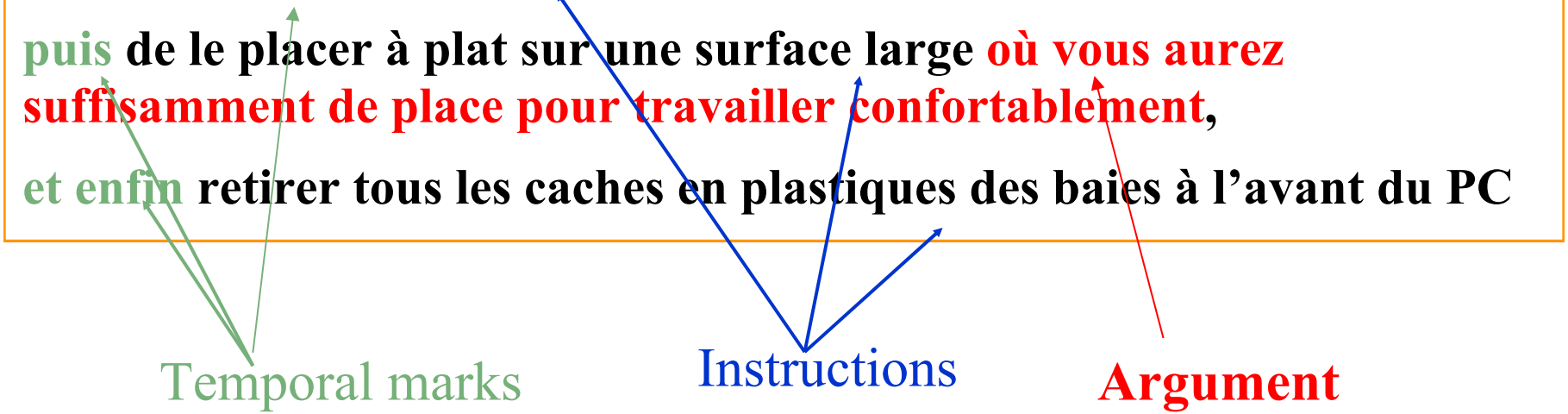
Imperative linear sequence → instruction < (temporal mark) < imperative linear sequence / instruction

Instruction → (iterative expression), action, (goal) (argument)+, (reference), (picture)+, (warning)

La premiere étape consiste à ouvrir entièrement le boîtier,

puis de le placer à plat sur une surface large où vous aurez suffisamment de place pour travailler confortablement,

et enfin retirer tous les caches en plastiques des baies à l'avant du PC



Conclusion

- Pour les systèmes question-réponse, problème de la quantité et qualité des pages renvoyées par le moteur de recherche :
  - nécessité de filtrer les pages en fonction des types de questions

- Intérêts du Web pour les systèmes question-réponse :
  - étudier en détail la forme et le contenu des textes du Web pour :
    - identifier les types de données à manipuler,
    - comprendre et résoudre les problèmes auxquels un système est confronté
  - définir des méthodes qui élaborent une réponse appropriée en prenant en compte des critères mis en évidence lors de l'étude de corpus