Les corpus web et l'approche textométrique :

conditions de collecte et nécessaire combinaison d'approches quantitatives et qualitatives

Valérie Beaudouin
LTCI – Télécom ParisTech

Séminaire CEDITEC
21 mars 2014

Principes de la lexicométrie

=> explorer le social à travers la médiation du texte

- Corpus, ensemble de textes caractérisés par :
  - Les conditions de production : date, genre…
  - Les caractéristiques des émetteurs : qui parle ? D'où ?

  => contrôler les variations du contexte d'énonciation
- Méthodes
  - Approche supervisée (hypothèse préalable)
    - Décrire des classes de textes définies a priori
    - Affecter des textes à des classes prédéfinies
  - Approche non supervisée (inductive)
    - Construire des typologies à partir des textes
    - Identifier proximités et distances entre textes, entre mots
- Aides à l'interprétation
  - Visualisation
  - Listes

Lexicométrie

- Construction d'un tableau lexical

	mot 1	...	mot j	...	mot n
Unité textuelle 1	0		1		1
...					
Unité textuelle i	1		0		0
...					
Unité textuelle k	1		1		1

- **Variables** : formes graphiques vs mots lemmatisés ; mots grammaticaux conservés ou éliminés, locutions et sigles considérés ou non comme une seule variable
- **Intersection** : fréquence du mot dans l'unité textuelle / présence-absence
- **Traitements statistiques** :
  - Analyse factorielle des correspondances, classification ascendante hiérarchique ou descendante hiérarchique....
  - Spécificités, richesse, évolution…

Les méthodes de statistique textuelle

Double origine (années 70) :
- Travaux en lexicométrie
- Travaux en analyse des données

Analyse exploratoire
- Spécificités, richesse…
- Analyse factorielle-classification
- Distance inter-textuelle

Corpus de textes « générés »
- enquêtes/entretiens

Mutations

Nouveaux corpus

Corpus issus du Web
- Passage à l'échelle
- Discours produits en situation « naturelle »
- Organisé par genres structurants
- Imbriqués dans des plateformes
- Structure réticulaire, multimedia, dynamique

Nouvelles méthodes

Apprentissage automatique
- Méthodes supervisées
- Méthodes non supervisées

Analyse des réseaux sociaux

CORPUS DU WEB

Les corpus liés au Web

- Le réseau ressource pour la constitution de corpus de textes
  - réservoir exceptionnel de textes numérisés
  - Illusion de l'exhaustivité et de l'accessibilité

- Comment constituer un corpus?
- Comment gérer le passage à l'échelle ?

- On ne peut réduire des documents web à du texte
  - Multimedia (-> comment articuler le multimedia dans l'analyse?)
  - Interactif (-> social network analysis)
  - Dynamique temporelle (-> comment la traiter?)

Les corpus du web

- Reconstituer la structure hiérarchique, réticulaire
  - Explorer, crawler en profondeur
  - Extraire les contenus pertinents
  - Structurer en base de données
- Mettre en relation les textes avec les profils de leurs auteurs
- Préserver les liens entre textes
- Conserver les documents liés (images, vidéo) et leurs annotations.

De la nécessité de croiser les regards

- Varier la constitution du corpus
- Tester différents outils
- Croiser avec d'autres méthodes
  - Analyse qualitative
  - Ethnographie
  - Sémiologie
  - Economie

EX 1 : HYPERTEXTE

- Pages personnelles (les ancêtres des blogs) :
  - Objet multimedia complexe
  - Ne peut être réduit à du texte
- Construire un corpus : deux approches
  - Les pages personnelles d'un réseau d'individus connectés
  - Les pages perso visitées par un panel d'internautes
- Construction d'une typologie
  - Traits de vocabulaire
  - Traits hypertextuels (liens, images, traits html)...

Construire une typologie des pages personnelles

- 100 000 pages vues par un panel d'internautes (panel NetValue-Nielsen//Netratings)

- Pages décrites
  - par des traits hypertextuels (liens, images, traits méta)
  - par le vocabulaire (taille du vocabulaire et place des pronoms)

- Une chaine de traitement complexe pour passer de la page à sa description par des traits
  - projet RNRT SensNet : FT, NetRatings, LIMSI, PIII
  - Beaudouin V., Fleury S., Pasquier M., Habert B. & Licoppe C. (2002). "Décrire la toile pour mieux comprendre les parcours. Sites personnels et sites marchands", Réseaux, Vol. 20, n°116, p. p. 19-51.

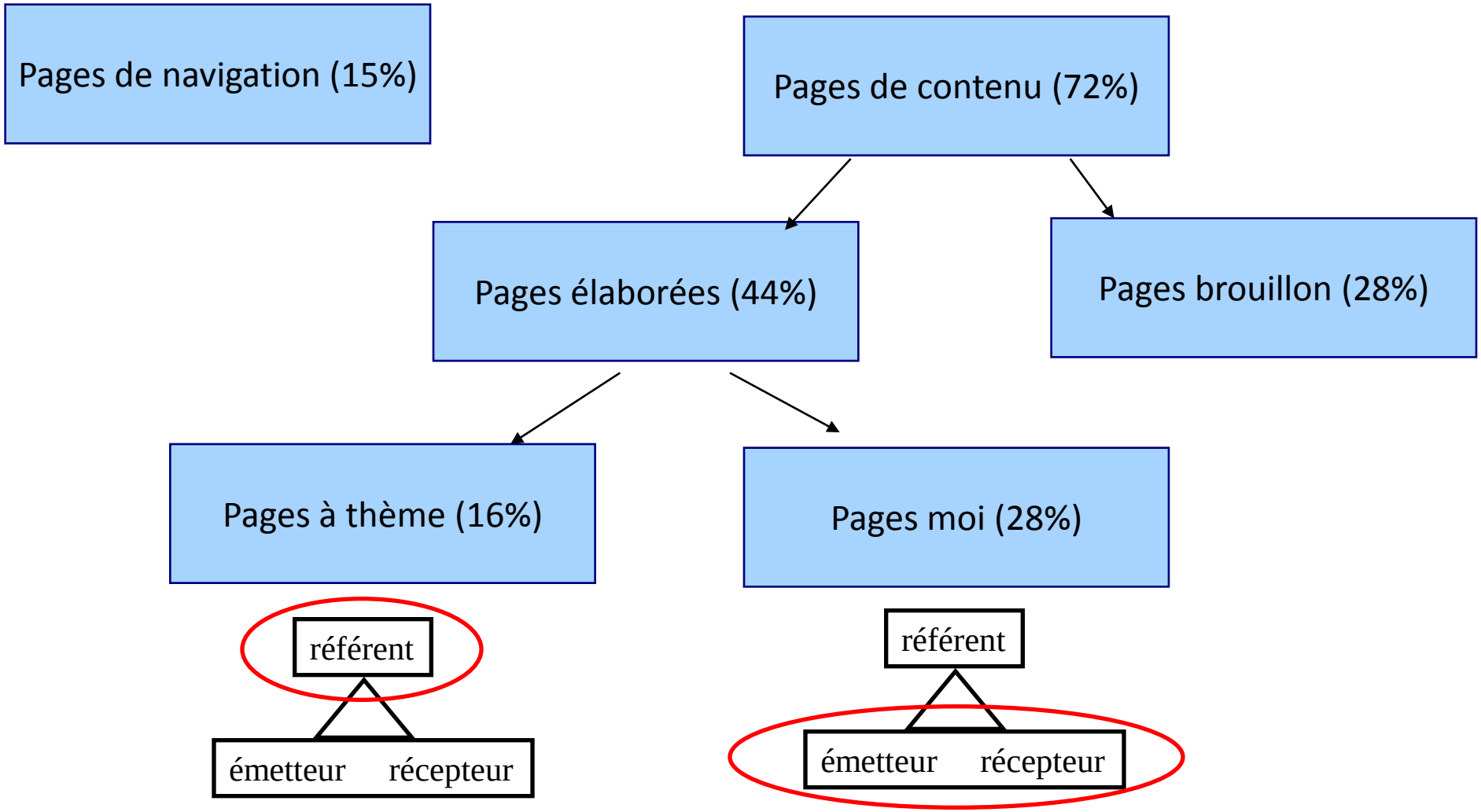
Les liens hypermedia selon le serveur d'hébergement

Serveur d'hébergement	Nb moyen de pages visitéees par site	Nb moyen de liens internes	Nb moyen de liens externes	Nb moyen d'images	Nb moyen d'images externes	Nb moyen de liens vers BAL
free_fr	6,4	20,1	3,5	7,0	0,8	0,4
perso.wanadoo.fr	5,4	9,6	1,9	5,1	0,3	0,5
perso.club-interne	5,3	13,8	3,6	7,1	1,4	0,7
www.chez.com	5,0	19,0	4,2	5,7	0,6	0,9
le-village.ifrance.c	4,8	7,0	3,1	5,0	0,6	0,8
www.multimania.c	4,8	6,1	3,7	4,6	0,5	0,4
ifrance.com	3,9	5,7	3,5	5,8	0,5	0,5
www.geocities.com	3,0	6,5	6,5	6,7	3,4	0,4
autres	3,2	10,3	4,5	6,5	0,9	0,5

Rq : sites perso.wanadoo visitées 6,5 pages en moyenne, 1,9 liens externes, 7 liens internes

Une très faible part des sites est effectivement visitée

Typologie des pages personnelles



Pages visitées : types de pages

- Les types de pages sont caractérisés par des configurations spécifiques de traits.
- Les types peuvent être interprétés comme des états différents d'élaboration de la page
  - Pages sophistiquées: beaucoup de liens, beaucoup d'images
    - pages d'experts avec maîtrise de l'écriture html, sans pronoms
    - pages avec beaucoup de pronoms personnels, liens vers boîte aux lettres
  - Pages "brouillon" : peu de liens, peu d'images, peu de pronoms, peu de maîtrise dans la définition de la forme de la page
- On mesure plutôt le degré d'élaboration de la page que des sous-genres au sein des sites personnels.

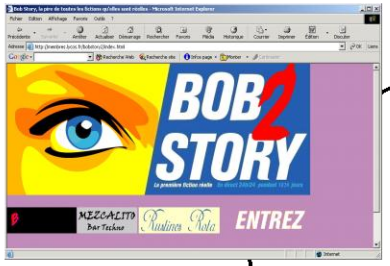
TEMPS

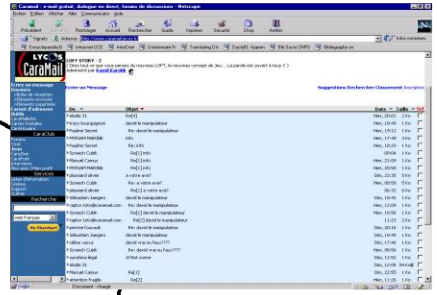
Constitution des publics autour du Loft

- Loft Story: un objet paradoxal
  - Des médias critiques
  - Un public enthousiaste : Une usine à bavardage
- Enquête sur les publics qui se constituent en lien avec des événements médiatiques
- Objectifs
  - Entrelacement des pratiques
    - Comment les pratiques des téléspectateurs articulent les différents médias et modes d'interactions (télévision, internet, presse papier, radio, téléphone…).
  - Comment les médias de masse rythment et informent les pratiques des spectateurs
  - Isoler le rôle spécifique du média Internet dans la constitution des publics de l 'émission

Loft Story agit comme une « usine à bavardage »

Regarder/
écouter

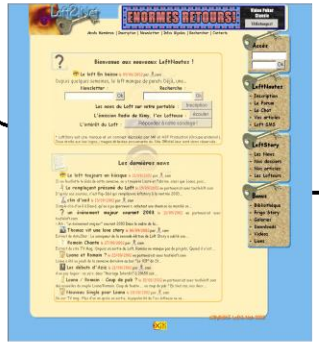


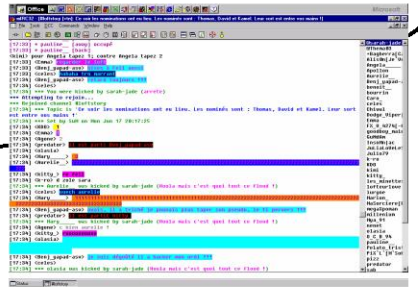


Internet, un relais idéal du Loft :

- Lieu de recherche d'information
- Lieu d'échanges interpersonnels

Rechercher sur le Web





En parler dans les chats et les forums

➔ **Quelle place pour Internet dans la constitution des publics de l'émission?**

Méthodologies articulées (1)

- Analyse des traces de connexions Internet au Loft 1 - 2001 (Projet coopératif RNRT – FTR&D, NetValue, Limsi, Paris III)
  - Identification des sites, des requêtes moteurs, des canaux de chat liés à l'émission
  - Analyse des parcours internet liés à l'émission
  - Profil des loftiens internautes

Méthodologies articulées (2)

- Vingt-quatre entretiens approfondis avec des spectateurs de Loft Story 2 sur l'ensemble de leurs pratiques liées à l'émission

- Analyse des interactions dans les chats
  - 3 semaines d'enregistrement continu sur deux canaux : M6 et Voila
    - Analyse conversationnelle des extraits d'interactions
    - Textométrie sur corpus de chat

Quand la télévision reconfigure les usages d'Internet

- Transformation des pratiques de consultation et d'interactions sur Internet
  - **Au niveau temporel**
  - **Au niveau thématique**
    - Le loft comme sujet de discussion (en face à face, par téléphone, mail, sur les chats…)
    - Le loft comme thème de recherche (dans les autres médias, sur internet, dans l'entourage)
  - **Au niveau formel (syntaxique)**
    - Les échanges dans les chats importent des formats d'échange propres à l'émission (vote, quizz)
    - Ces formats engagent des collectifs plus importants que les échanges habituels sur les chats

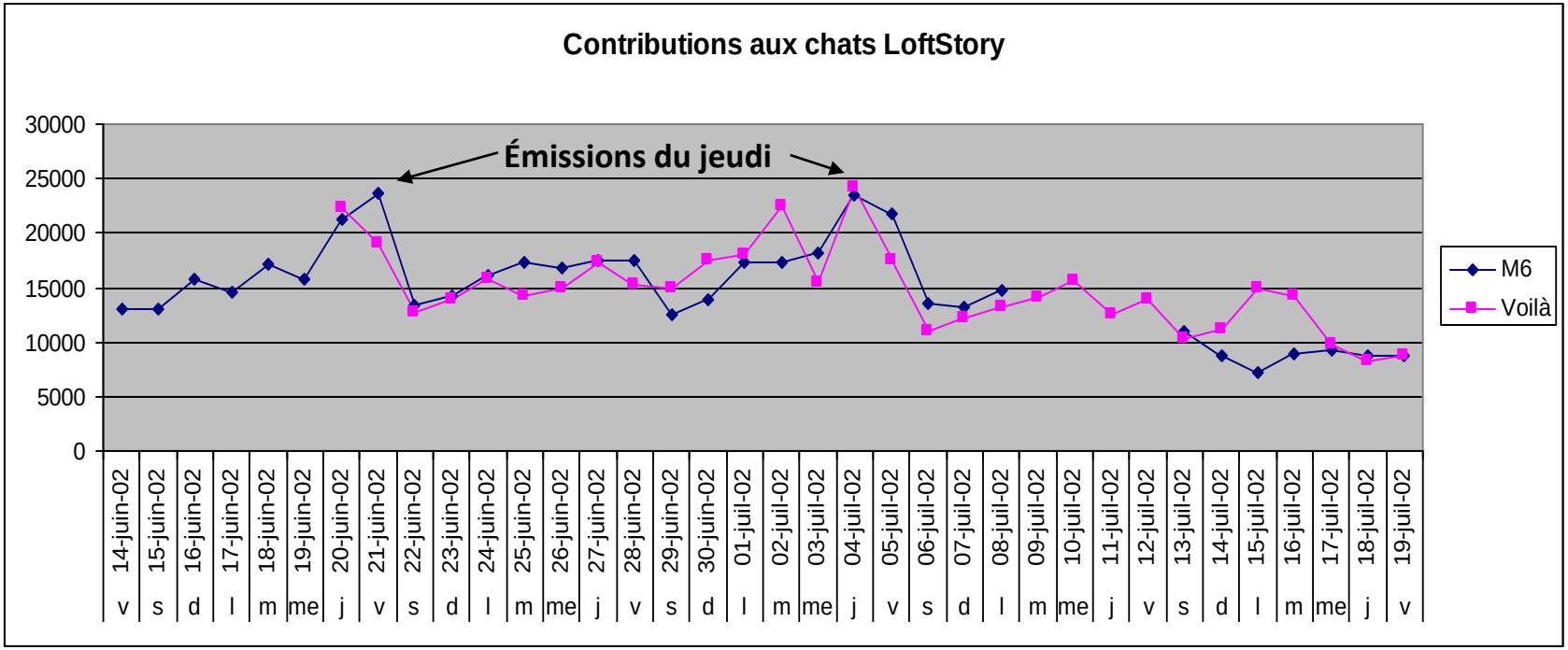
1) L’émission comme cadre temporel

- La fréquentation de sites liés au Loft suit le rythme de l’émission
  - Trois semaines après le début de l’émission, les trois-quarts des 573 internautes se sont connectés à un site « Loft » => effet rapide de l’événement
  - Pics de connexion sur les sites Loft le lendemain de la grande émission du jeudi soir
- Des groupes se constituent (sur Internet) en phase avec l’émission

- Le rythme et l'intensité des échanges suit de près le rythme des émissions M6

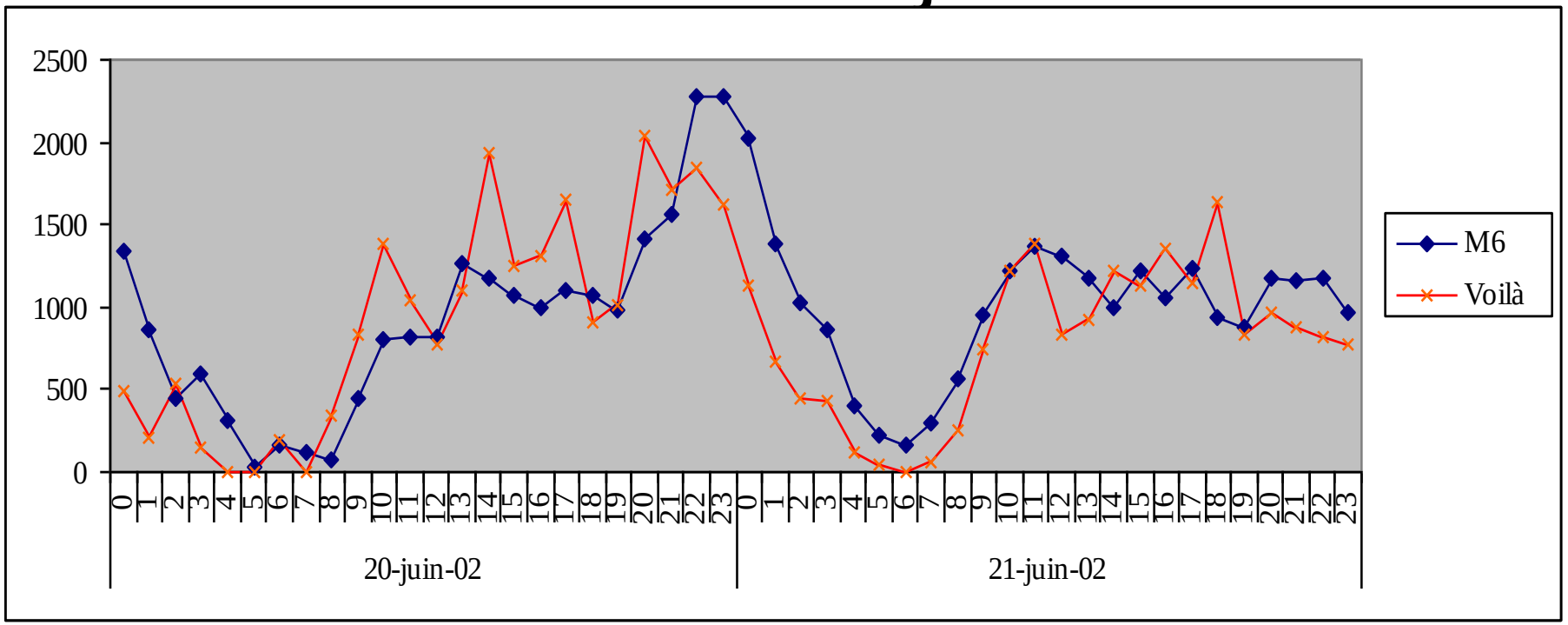
Session Start: Sat Jun 29 00:00:01 2002
Session Ident: #LoftStory
[00:00] <le_chat_F> si c'�tait dans le journal tu le saurais pas car tu ne sais lire que pif magazine
[00:00] <__Anti_David_> Bonjour j'sui Benjamin Castaldi d�guis� chut fo pa le dire;)
[00:00] * iowa has quit IRC (Read error: Connection reset by peer)
[00:00] <le_chat_F> c de ton niveau
[00:00] <derf33> M� bien sur
[00:00] <le_chat_F> maintenant tu me lache
[00:00] * myrtille- has quit IRC (Quit: Voila)
[00:00] * dsl_G_RESON has quit IRC (Quit: using sirc version 2.211+KSIRC/1.2.1)
[00:00] * Voyou has quit IRC (Quit: Voila)
[00:00] <Boucheron> loana r�vise tes vannes de laurent ruquier
[00:00] <Loana_du_Loft_[La_Vraie]> non le chat F c toi ka commenc�
[00:00] <Loana_du_Loft_[La_Vraie]> mintenan tu d�gustes
[00:00] <le_chat_F> non
[00:01] <le_chat_F> c pas moi qui d�guste cocotte c'est toi
[00:01] <__Anti_David_> Dehorsssssss Davidddddddd
[00:01] <filouv> bon vous avez po fini oui
[00:01] <filouv> on va pas passer la soiree avec ca
[00:01] <Boucheron> on dirait une enfante
[00:01] <Loana_du_Loft_[La_Vraie]> @filou ils sont d�chain�s
[00:01] <__Anti_David_> Moi je propose k'a la sortie du Loft on aille o Studio jeter des dicos sur la gueule de DAVID et d'ANGELA!!!!!

Le rythme des chats suit les grands moments de l'émission



- Le nombre de contributions sur les chats M6 et Voilà est globalement identique
- L'activité sur les chats M6 et Voilà suit le rythme de l'émission :
  - Forte augmentation de l'activité autour de l'émission du jeudi soir
  - Déclin progressif de l'activité après la clôture de l'émission

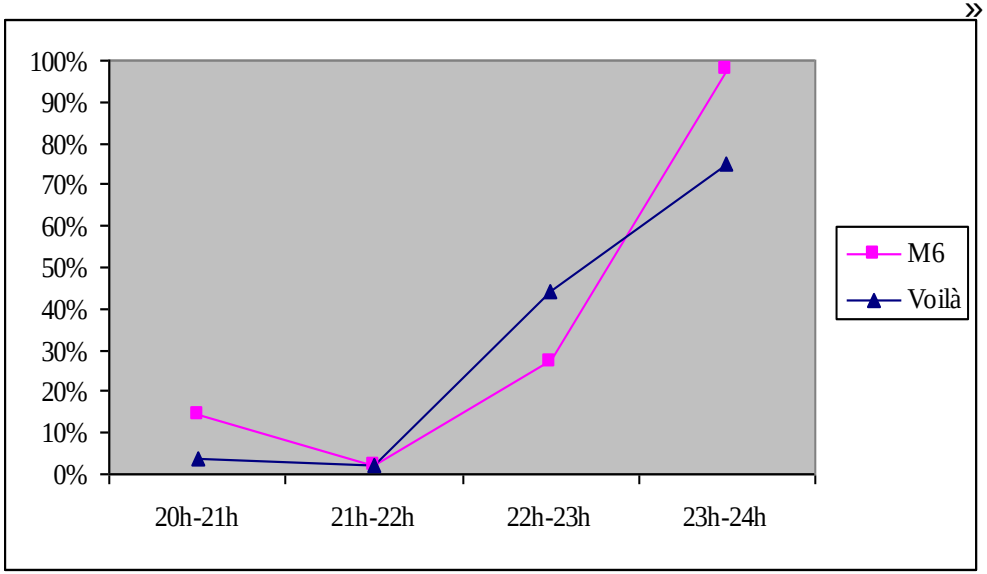
Emission du 20 juin 2002



- Sortie de Kamel, alors que David était le plus bas dans les votes tout au long de l'émission
- Pic d'activité au moment de l'annonce de la sortie de Kamel, mobilisation plus intense sur M6 que sur Voila

Part des messages critiques

- Identification des messages critiques par rapport à M6



» Vocabulaire spécifique : huissier+(37), truqu+er(77), hont+e(69), felic+(215), m6(136), magouille+(54), mauvais+(43), trich+er(42), boire.(25), castald+(37), prod+(48), chaine+(22), femme+(27), degout+er(25), denonc+er(15), perdre.(29), arnaque+(20), boouh(15), boouuhh(15), boycott+(17), dody(14), perdant+(15), zw(14), decu+(13), expressi+f(8), prochain+(14), producti+f(8), suisse+(10), vainqueur+(9), jaune+(12), france(23), chat+(35), colere+(10), debut+(14), facon+(12), gens(36), liberte+(8), merde+(39), plateau+(13), preuve+(9), truc+(19), yeux(12), pa+yer(16), prevoir.(10), regard+er(53), fait(109), possi+ble(18), travail<(19), boycot+(13), castelli(8), encule+(8), filliale(9), hyppo(14), kick+(23), laur+(11), lofteuse91(9), mariesolange(10), mumu(16), pf(8), pognon+(11), polio+(9), sylvano(15), tele(18), tiny(10), comique+(7), droit+(16), fina+l(32), hilare+(7), integre+(5), italien+(5), parti+(19), plein+(15), premier+(12), vive+(81), audience(5), avion+(7), bebe+(7);

- Part des messages critiques pendant la soirée du 20 juin
- Sur le chat M6, ces messages occupent l'intégralité des échanges à de 22h45, sur Voila le phénomène est un peu moins intense.

PLATEFORME

La critique amateur sur Allociné

- La place des amateurs dans la critique de cinéma
- L’organisation du monde des amateurs

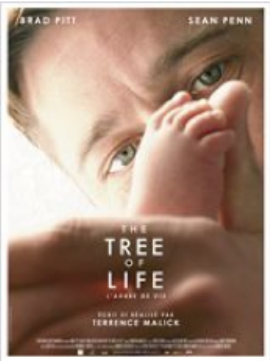
L'enquête

- Entretiens avec des professionnels d'AlloCiné
- Analyse sémiologique de la plateforme
- Enquête par questionnaire et entretiens sur les utilisations du site par des 15/20 ans
- Extraction des données de la plateforme pour 140 films (sortis en 2011 et encore à l'affiche) :
  - Caractéristiques du film
  - Critiques pro
  - Critiques amateurs
  - Profils des rédacteurs

Page de film

The Tree of Life

Séances | Bandes-annonces | Casting | Critiques | Photos | DVD, VOD | Le saviez-vous ?



Date de sortie **17 mai 2011** (2h 18min)
Réalisé par Terrence Malick
Avec Brad Pitt, Jessica Chastain, Sean Penn › plus
Genre Drame, Fantastique
Nationalité Américain

Presse 3,5
Spectateurs 2,7

Voir la bande-annonce

Page de critique

brianpatrick
20 abonnés | Lire ses 666 critiques | Suivre

5 - Chef d'oeuvre

C'est la Genèse selon Terrence Malick. Monsieur Terrence Malick doit se poser des questions quant à son verre à trois quart vidé. Ce film ressemble fortement à l'Odyssée de l'espace de Kubrick, le même genre. Certes sur arte il y a énormément de film d'art et d'essai filmés par des bobos, mais tous ratés ou complètement nuls. Par contre il faut reconnaître que Terrence Malick et Kubrick sont des maîtres de genre, et certes j'adhère complètement. Donc du génie, un magnifique film.

Ajoutée le 15 juin 2011 à 21h59

J'aime Signaler un abus

Page de membre

Suivre | Discuter

brianpatrick (AlloCinéen)

Membre depuis 2384 jours
0 point
Voir son profil complet

Ses Critiques (666)

Tri : Les plus récents 1 / 67 1 - 10 sur 666 résultats

Jardins secrets
Avec Daniël Boissevain, Cystine Carreon
Série néerlandaise - Comédie dramatique
Voir les photos | Diffusions

Sa note : (2)
Sa critique : C'est une série sans grand intérêt, il ne se passe pas grand chose, une série d'ailleurs diffusée à 2h00 du matin.
Ecrite le 29 mai à 12h30 - Signaler un abus - Permalien

Cowboy
De Benoît Mariage
Avec Benoît Poelvoorde, Gilbert Melki
Film français - Comédie
Bande-annonce

Extraction, aspiration et reconstitution d'une base de données à partir d'Allociné (réalisation technique Denis Vilar)

Films

Titre, réalisateurs, comédiens, date sortie, note presse, amateur

Critiques presse

Titre film, notes, extraits de critiques, nom du critique, support, date

Critiques amateurs

Titre film, **pseudo**, note, critique, date

Membres

Pseudo, club300, nb de critiques, nb d'abonnés, **site monallocine**…

Profils membres

Infos présentes sur le **site monallociné** : amis, fan de

Sur 140 films sortis en 2011, à l'affiche en fin d'année

40 000 critiques ont été rédigées par

18 000 membres d'Allociné

dont 10 000 ont indiqué au moins un élément dans leur fiche profil

(Quatre fois plus de notes que de critiques)

Club 300

- Activité critique plus intense :
- Notation plus mesurée (mediane =3 au lieu de 4)
- Délai de publication : 45% de leurs critiques publiées la semaine de la sortie (vs 33%)
- Choix de films : préférence pour films valorisés par la critique presse
- Longueur des critiques : 32% des critiques du Club 300 ont plus de 200 mots (vs 8%)
- Types de critiques : 80% des critiques du Club300 de type 1:
  - centrées sur le film, évitant le je-vous, la recommandation et l'expression émotionnelle
  - référence au champ du cinéma dans toutes ses dimensions (réalisateurs, comédiens, festivals, critique…)

Les deux modalités de l'exercice critique

- Deux formes de la critique profane :
  - habitués-cinéphiles *vs* les novices-simples spectateurs (Allard, 2000)
  - High art *vs* popular aesthetic discourse (Van Venrooij 2010; Verboord, 2013)
- Sur notre corpus (échantillon de 10 000 crit.)

Critiques centrées sur le film argumentation 48%

Critiques centrées sur la réception émotion 52%

1 : Critiques centrées sur le film (48%)	2 : Critiques centrées sur la réception (52%)
3è personne	Je/vous
Genres : Drame / comédie dramatique / biopic	**Genres** : comédie, animation / action/ aventure / espionnage / horreur / épouvante / thriller
Mimesis : sujet de l'histoire, personnages	**Émotions** : super, adorer, décevoir, sympa, ennuyer,
Forme : réalisation, jeu, interprétation.	**Prescription** : conseiller, allez-y…
Références à l'univers du cinéma (histoire, œuvres, prix, techniques…)	
Mots typiques : réalisateur, vie, mise en scène, homme, spectateur, femme, sujet, cinéma, personnage, plan, montrer, sentir, dernier, filmer, sembler, amour, force, monde, vivre, prendre, image, jeune, côté, cinéaste, grand, jour, beauté, politique, devenir, humain, regard, rendre, travail, laisser, réalité, parler, lumière, comprendre, œuvre, interprétation,	**Mots typiques** : sympa, super, rire, adorer, dessin animé, gag, rigoler, drôle, marrant, décevoir, action, divertissement, animation, annonce, moment, ennuyer, humour, comédie, saga, détente, divertissant, navet, requin, fan, nain, vivement, chat, recommander, conseil, vraiment, destination, spécial, sympathique, bande, voir, bond, top, agréable…

1 : Critiques centrées sur le film (48%)	2 : Critiques centrées sur la réception (52%)
« Les esprits cartésiens et scientifiques éprouvent d'évidence les pires difficultés à pénétrer le cinéma de Bruno Dumont, qui effectivement exige qu'on abandonne à l'entrée de la salle ses certitudes. Voir se produire à l'écran des événements surnaturels – un exorcisme, un arrêt d'incendie et une ressuscitation – ne peut en rien justifier la disqualification d'un film. L'intérêt se situe ici sur la forme : comment le réalisateur s'y prend t-il pour mettre en scène cette histoire absolutiste, faite de silences et d'une sécheresse qui peut rebuter » [suivent encore 20 lignes] (Hors Satan, Drame, note_3, gros contributeur)	*« C'est frais, drôle et surtout sans prétention. J'ai passé un excellent moment et je le conseille vivement. »* (La Fée, comédie, note :4, novice) *« les gags sont prévisibles mais ce film n'est pas si mauvais je trouve, même si les meilleures scènes comiques sont dans la bande annonce »* (On ne choisit pas, Comédie, Aventure, note : 3,5, contributeur moyen)

Du novice …

Une seule critique

- Note
  - moyenne de 3,63
  - médiane de 4,5
- Genres : comédie, action
- Longueur
  - moyenne : 52
  - médiane : 30
- 70% critiques centrées sur réception – 30% sur le film

à l'habitué

> plus de 50 critiques

- Note
  - moyenne de 3,30
  - médiane de 3,5
- Genres : drame, comédie dramatique
- Longueur
  - moyenne : 115
  - médiane : 73
- 40% centrées sur la réception – 60% sur le film

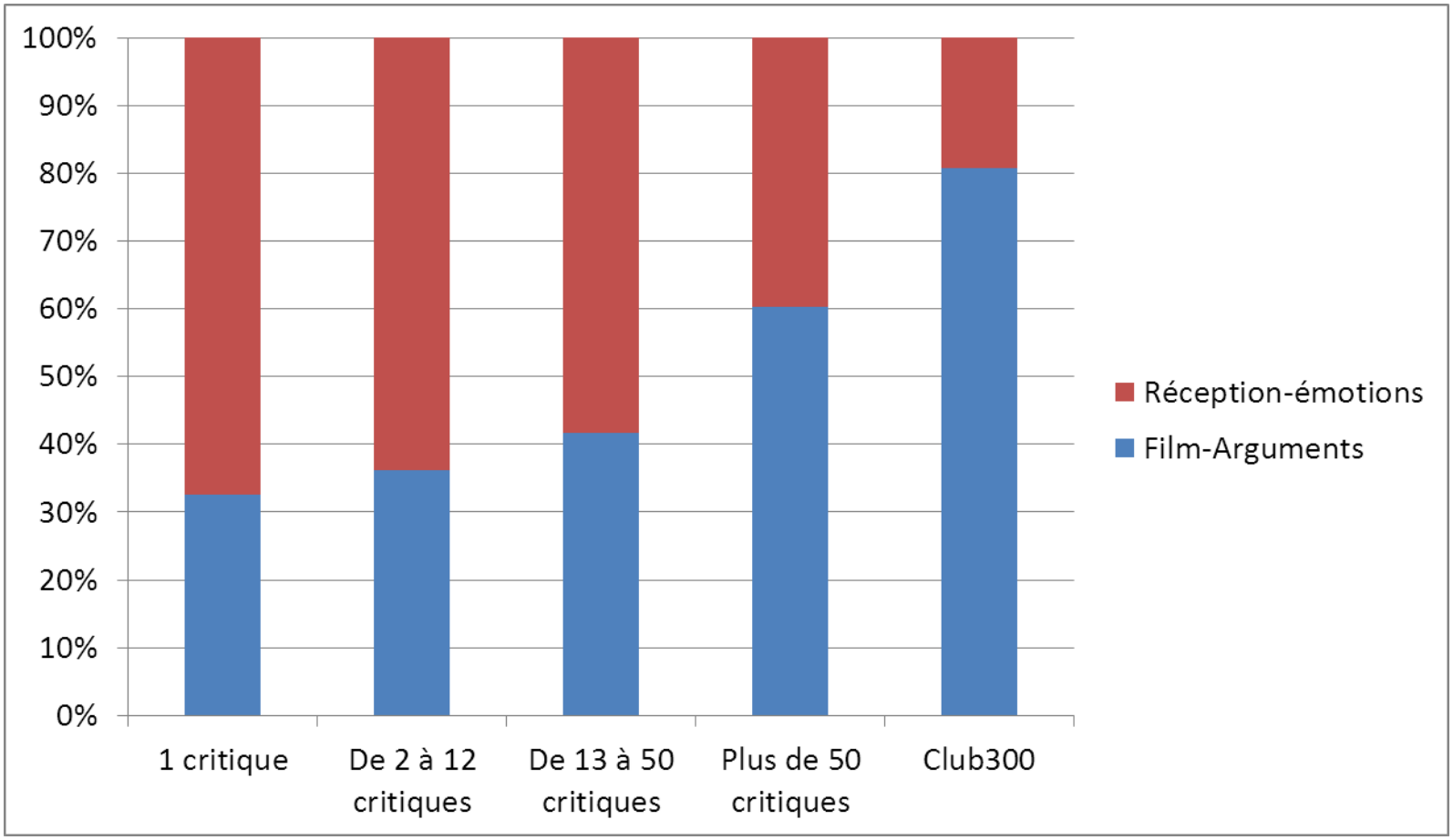
Le Club 300 : élite de la critique amateur

- Créé par Allociné en 2008
- Critiques amateurs internes ou externes avec pouvoir d'influence
- Rétributions symboliques
- Echanges de procédés réputationnels

Le Club 300 en activité

- **Activité** critique plus intense (14 en moy)
- **Notes** plus mesurées (moy=med=3)
- **Délai** de publication (45% semaine de sortie vs 33%)
- **Genres** : drame+, films d'action+, thrillers+, comédie--
- **Films** : plutôt films d'auteurs que film commercial
- **Longueur** : 32% des critiques du Club 300 ont plus de 200 mots contre 8% pour les autres
- **Type** : 80% centrées sur le film, 20% sur réception

=> Proximité avec critique professionnelle


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1 critique
De 2 à 12 critiques
De 13 à 50 critiques
Plus de 50 critiques
Club300
Réception-émotions
Film-Arguments

ANALYSE DES RÉSEAUX SOCIAUX

Réseaux sociaux

- Déplacement :

  Caractéristiques des individus / textes

  – > Liens entre individus / textes
- Relations, liens
- Structure des liens

- Question : comment articuler analyse de discours et analyse de réseau ?

Réseau et trajectoires des écrivains sur le Web

- Contexte :
  - Présence relativement faible des écrivains sur le Web
  - Emergence de nouveaux auteurs nés dans le numérique
- Deux postures d'auteurs face au web : lieu d'écriture / vitrine
- Construire le monde de l'écriture numérique (revue, maison d'édition, réseau…)
- Tension dans les trajectoires :
  - Individu vs collectif
  - Intégré vs éclaté

Comprendre la logique des trajectoires d'écriture sur le Web

Cartographier la littérature française contemporaine sur le Web

- Un point d'entrée : remue.net
- Exploration de proche en proche via les liens intersites (navicrawler)
- Qualification des nœuds en fonction des degrés entrants et sortants (hubs, authorities)

Reconstituer les trajectoires d'écriture

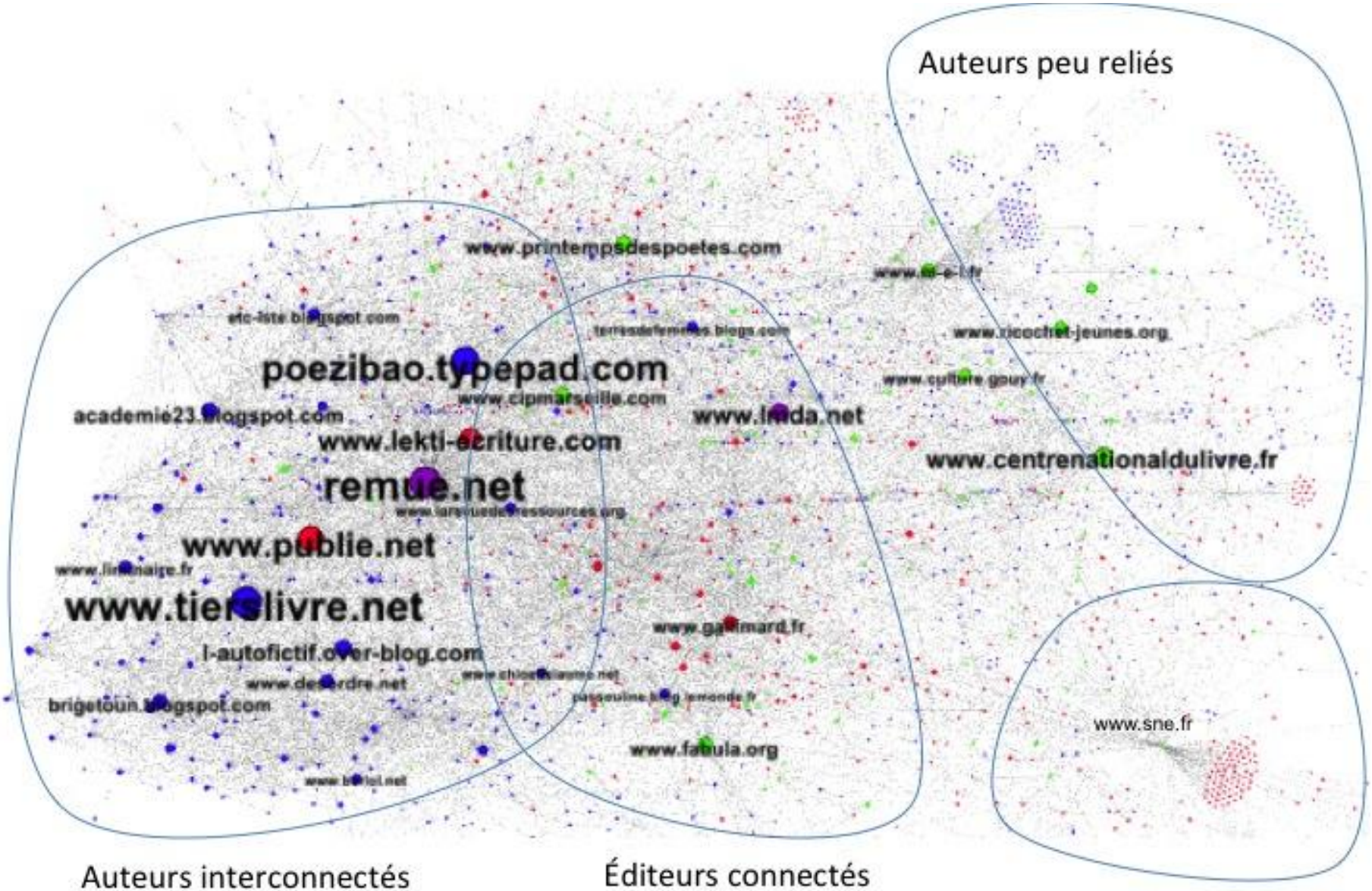
- Sélection de sites saillants sur le graphe
- Exploration des lieux de publication connexes
- Reconstitution de la trajectoire
  - via l'exploration des archives du Web
  - via des entretiens

Comment les activités d'écriture et de promotion de soi se trouvent projetées dans le Web
Comment les pratiques d'écriture se déplacent au fil des innovations sociotechniques (site perso -> blog-> CMS…) ?
Comment émergent de nouveaux genres d'écrits ?

Fragilité des carrières de publication sur le Web

- Le numérique : l'illusion de la publication ouverte à tous
  - Facile
  - Gratuit
  - Sans filtre
- Ou le principe de réalité
  - Abandons rapides faute d'audience
  - Abandons malgré le succès
    - Innovations socio-techniques qui demandent de nouvelles compétences
    - Abandon au profit de métiers du Web (vers professionnalisation)
    - Lourdeur de l'engagement

  Exemples :
  - 8 abandons sur 14 sites à forte audience en 1998 :
  - 3 des 20 individus les plus visibles aujourd'hui sur le Web ont ancienneté antérieure à 2003.


Auteurs peu reliés
www.printempsdespoetes.com
poezibao.typepad.com
www.cipmarseille.com
academie23.blogspot.com
www.lekti-ecriture.com
remue.net
www.publie.net
www.tierslivre.net
www.linda.net
www.centrenationaldulivre.fr
www.ricochet-jeunes.org
www.culture.gouv.fr
www.gallimard.fr
www.fabula.org
www.sne.fr
Auteurs interconnectés
Éditeurs connectés
Éditeurs peu reliés

Le web pour les écrivains

- Territoire d'écriture peu investi
  - un tiers des écrivains ont leur site fin 2008
- Parmi ceux qui ont leur espace de publication numérique, deux profils très différents

Site = vitrine de promotion

Auteurs très peu reliés à d'autres
Auteurs édités par maisons d'édition « classiques »
Cités par institutions comme la Maison des écrivains, le CNL…

Site = **lieu de création** et de promotion
Auteurs très interconnectés, reliés entre eux, à revue, à maison d'édition

Auteurs actifs sur les réseaux sociaux

Lieu d'expérimentation littéraire dans le numérique

Auteurs publiés « papier » et auteurs exclusivement numériques

François Bon : Remue.net, le tiers livre et Publie.net

Individuel

François Bon : site personnel (1997-2000)

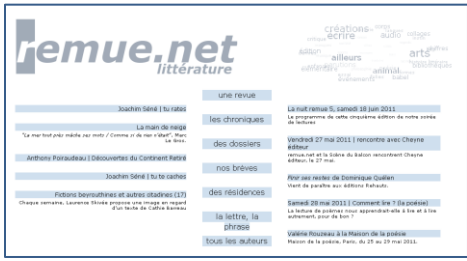
le tiers livre : blog

(2005-aujourd'hui)

1997 2000 2003 2005 2007 2011

Collectif

Remue.net : collectif d'auteurs



(2000-aujourd'hui)

Publie.net : page perso, extension de remue.net et tierslivre.net (2002–2007)

Publie.net coopérative d'édition (2008–aujourd'hui)

MÉTHODES DU WEB – APPRENTISSAGE AUTOMATIQUE

Machine learning – passage à l'échelle

- Sous-domaine de l'intelligence artificielle
- Machine apprend à partir de données empiriques
  - Crée un modèle de comportement
- Deux catégories
  - Apprentissage supervisé
  - Apprentissage non supervisé
- Traits utilisés : mots, parties du discours…
- Algorithmes : SVM, Bayésien naïf…
- Mesure de la qualité : précision et rappel (F-score)

Exemple : analyse de thème

Domaines d’application

- Analyse de contenu
- Analyse d’opinion (Sentiment Analysis and Opinion Mining)
- Attribution d’auteur

- => adapter les outils pour un usage en sociologie
- => monter en compétences dans le domaine

BILAN, PERSPECTIVES

Bilan sur les corpus tirés du web

- Les espaces d'interactions adaptés à la stat textuelle
  - espaces normés (netiquette)
  - règles locales de fonctionnement
  - formats d' intervention contraints
  - Le texte domine
- Les espaces d'autopublication sont problématiques
  - Premier genre numérique, mais
  - Genre créatif, en transformation permanente
    - Contenu et Forme
  - Genre multimédia et non pas textuel
  - Variabilité sur l'émetteur, sur le public, sur l'objectif, sur le genre
- Et les sites de médias sociaux : nouvelle frontière pour la recherche

Les conditions d'un bon usage

- Au préalable formuler des questions de recherche
- Utiliser les outils à titre exploratoire pour faire émerger des hypothèses
- Enrichir les textes de données contextuelles
  - Qui parle ? À quel public ? Dans quel contexte ?
- Croiser avec d'autres méthodes :
  - Enquête ethnographique, qualitative
  - Approche sémiologique
  - Tenir ensemble concepteurs de plateformes, contributeurs et utilisateurs

Bibliographie

- Référence
  - C. Barats, *Manuel d'analyse du Web*. Paris: Armand Colin, 2013.
  - C. Muller, *Étude de statistique lexicale. Le vocabulaire du théâtre de Pierre Corneille*. Paris: Larousse, 1967, réimpression aux éditions Slatkine, 1979, 1992, 1967, p. 382.
  - J.-P. et coll. Benzécri, *Pratique de l'analyse des données, Linguistique et lexicologie*. Paris: Dunod, 1981.
  - L. Lebart and A. Salem, *Statistique textuelle*. Paris: Dunod, 1994, p. 342 p.
  - M. Reinert, "Les 'mondes lexicaux' et leur 'logique' à travers l'analyse statistique d'un corpus de récits de cauchemars," *Langage et société*, no. n°66, pp. 5–39., 1993.
  - M. Reinert, "Classification descendante hiérarchique et analyse lexicale par contexte : application au corpus des poésies d'Arthur Rimbaud," *Bulletin de Méthodologie Sociologique*, no. n°13, p. xxx, 1987.
  - V. Beaudouin, "Statistique textuelle : une approche empirique du sens à base d'analyse distributionnelle," *Texto*, vol. V, 2000.
- Travaux /corpus cités
  - V. Beaudouin and D. Pasquier, "Organisation et hiérarchisation des mondes de la critique amateur cinéphile," *Réseaux*, vol. Les évalua, no. 183, 2014.
  - V. Beaudouin, "Trajectoires et réseau des écrivains sur le Web. Construction de la notoriété et du marché," *Réseaux*, vol. 175, no. 5, pp. 107–144, Nov. 2012.
  - V. Beaudouin, S. Fleury, and M. Pasquier, "Les pages personnelles comme terrain d'expérimentation," in *Les carnets du Cediscor*, vol. 8, F. Mourlhon-Dallies, F. Rakotonoelina, and S. Reboul-Touré, Eds. Presses Sorbonne nouvelle, 2004, pp. 143–164.
  - V. Beaudouin, T. Beauvisage, D. Cardon, and J. Velkovska, "L'entrelacement des médias dans la constitution des publics de Loft Story." France Télécom R&D, Issy-les-Moulineaux, p. 64 p., 2003.
  - V. Beaudouin and J. Velkovska, "Constitution d'un espace de communication sur Internet (Forums, pages personnelles, courrier électronique...)," *Réseaux*, vol. 17, no. 97, pp. 121–177, 1999.