



FOUILLE DE FORUM INTERNET

Forum Web Intelligence & Usages

Anna STAVRIANOU, Laboratoire ERIC
anna.stavrianou@univ-lyon2.fr

4 février 2010

Caractéristiques d'une discussion

■ **Sujet :**

«Y'a-t-il trop de commémorations en France?» (<http://www.liberation.fr/forums>)

■ **Internautes**

Ils s'identifient avec un *pseudonyme*.

■ **Messages**

Les internautes peuvent écrire un *nouveau* message ou *répondre* à un message existant.

■ **Relations**

Nous supposons que les relations «*répond à*» - tel message *répond* à tel message (un seul) - sont connues.

Y a-t-il trop de commémorations en France?

Vos commentaires

pipounette

▼ supprimer les fêtes religieuses

En effet, à part Noël, fête populaire à l'origine, pourquoi garder l'ascension, la Pentecôte, Pâques etc., fêtes catholiques qui n'ont plus lieu d'être si l'on compte les millions de musulmans et d'athées dont le total dépasse le nombre de catholiques pratiquants. On pourrait les remplacer par des fêtes laïques comme l'abolition de l'esclavage ou le jour où Chirac a dit NON à la guerre en Irak. En Grèce, il y a 2 fêtes nationales dont l'une commémorant le 28 octobre 1940, jour du NON, date à laquelle la Grèce a dit NON à Mussolini voulant emprunter le nord du territoire grec pour envahir la Russie.
Lundi 10 novembre à 11h30

Signaler au modérateur

Répondre



Slim

▼ Excellente idée

Remplacer une fête religieuse par une commémoration officielle des victimes de la Révolution française, 100% pour!!
Lundi 10 novembre à 12h20

Signaler au modérateur

Répondre



Jean Dubois

▼ Commémo

Et pourquoi-pas une journée de commémoration pour les fruits et légumes ?
Lundi 10 novembre à 16h06

Signaler au modérateur

Répondre



louis

▼ yessss!!!

Yessss! mais tout le monde a compris que pour notre bonne vieille république y a les bons et les mauvais morts...La révolution n'a rien changé à tout ça!
Lundi 10 novembre à 18h03

Signaler au modérateur

Répondre

Motivation

- Grand nombre et popularité de discussions en ligne
- Contenu intéressant
 - ❑ Opinions des internautes
 - ❑ Préférences, critiques de produits
 - ❑ Présentation d'idées politiques
- Nouveau domaine
 - ❑ Fouille et analyse de discussions
 - ❑ Faciliter la navigation
 - ❑ Faciliter l'extraction de connaissances

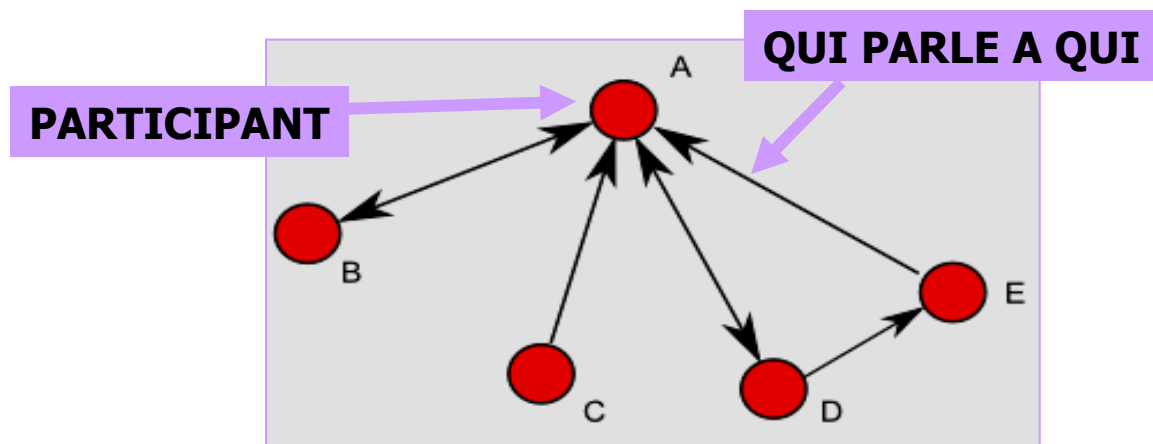


Plan

- **Méthodes existantes**
- Notre proposition
- Utilisation du modèle
 - Mesures
 - Recommandation
- Prototype
- Conclusion – Perspectives

Méthodes existantes

- Représentation
 - Réseau social des participants sous forme de graphes
- Objectifs
 - Analyser l'interaction entre les internautes
 - Identifier des communautés et des rôles



Méthodes existantes

- Nombreuses méthodes existantes

Zhang et al. (WWW, 2007)

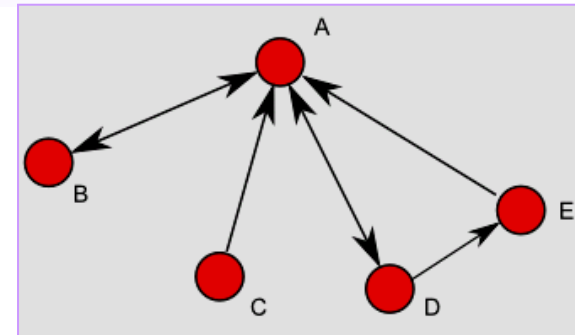
« *Expertise networks in online communities* »

- Motivation : analyse du forum de Sun (Java), identifier l'expertise des internautes
- Analyse : à combien de personnes on répond, nombre de réponses par rapport au nombre de questions, à qui on répond (répondre aux experts montre une expertise)

Fisher et al. (HICSS, 2006)

« *You are who you talk to: Detecting roles in Usenet newsgroups* »

- Motivation : interprétation de la structure d'une communauté afin d'identifier les rôles de chaque internaute
- Observer comment les utilisateurs interagissent entre eux (personnes centrales de la discussion, voisins etc.)



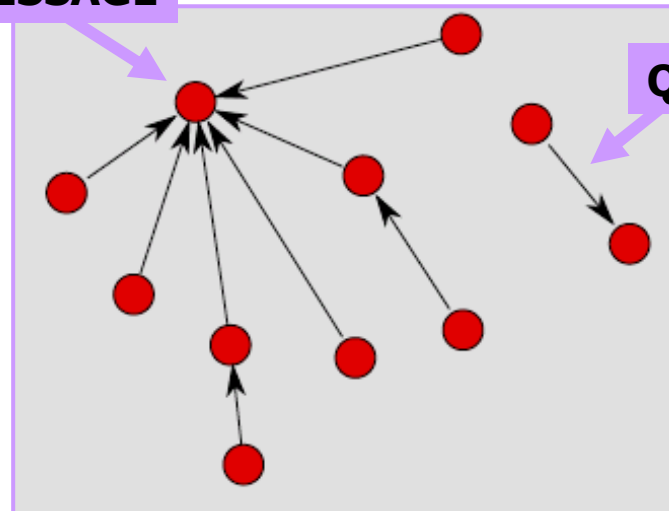
Limites des méthodes existantes

- Contenu, structure, opinion perdus
 - A quel moment un internaute a parlé ? Combien de fois?
 - Est-ce qu'il y a un conflit entre les internautes ?
 - Est-ce que les internautes ont tendance à discuter d'une manière négative/positive ?
- Point de vue limité, Réseaux Sociaux non adaptés
 - Répondre au contenu des messages et pas aux auteurs
 - Manque de relations entre les internautes

Notre modèle

- Représentation basée sur les graphes
 - **Sommets** : les *objets de type message*
 - **Liens** : relations «répondre à »

OBJET DE TYPE MESSAGE



QUEL MSG REPOND A QUOI

Post-Reply Opinion Graph (PROG)

Définition :

Nous représentons une discussion par un graphe orienté $G=(V, E)$

- V : l'ensemble des sommets des « *objets de type message* » qui ont le format

$$v = (m_v, op_v, u_v, tm_v),$$

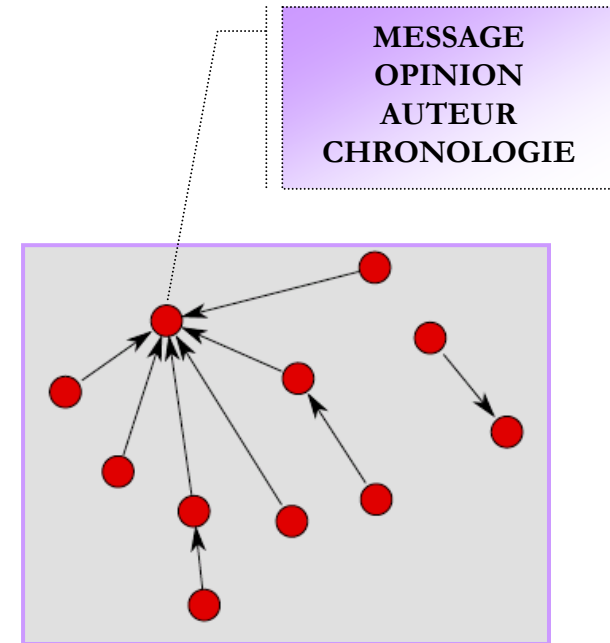
m_v est le contenu actuel du message

op_v est la polarité de l'opinion qui est incluse dans le message

u_v est l'internaute qui a écrit le message

tm_v montre la chronologie à laquelle ce message a été enregistré dans la discussion

- E : l'ensemble des liens où chaque lien $e_{ij} = (v_i, v_j)$ représente une réponse dirigée de v_i à v_j

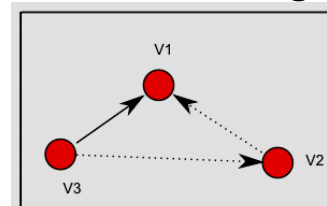
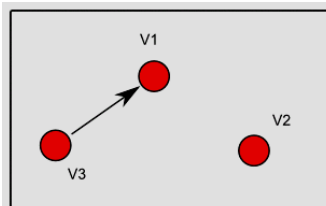
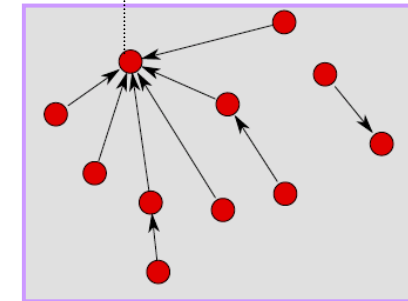


Caractéristiques des objets de type message

$$v = (m_v, op_v, u_v, tm_v)$$

- **Contenu du message** m_v
- **Opinion** op_v
 - Polarité de l'opinion : positive, négative, objective
 - Domaine d'Opinion Mining
- **Auteur** u_v
 - Possibilité d'extraire le réseau social du PROG
- **Chronologie** :
 - A travers les liens / tm_v – enrichir le graphe

MESSAGE
OPINION
AUTEUR
CHRONOLOGIE

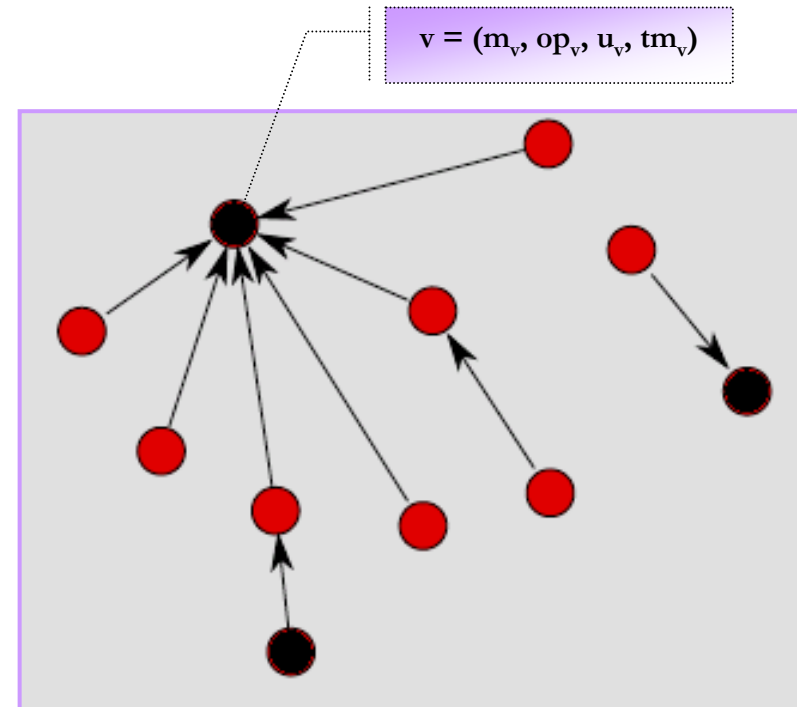


Plan

- Méthodes existantes
- Notre proposition
- Utilisation du modèle
 - **Mesures**
 - Recommandation
- Prototype
- Conclusion – Perspectives

Mesures basées sur l'opinion

- **Opinion (Opinion Mining)**
 - **Statut d'un internaute**
$$avgOpFromUser(u) = \frac{\sum_t op_{v_t}}{|msgs(u)|}$$
 - Statut vers un internaute
 - Réactions vers un message
 - Sommets d'une polarité
 - Opinion Moyenne
 - Variété de l'opinion



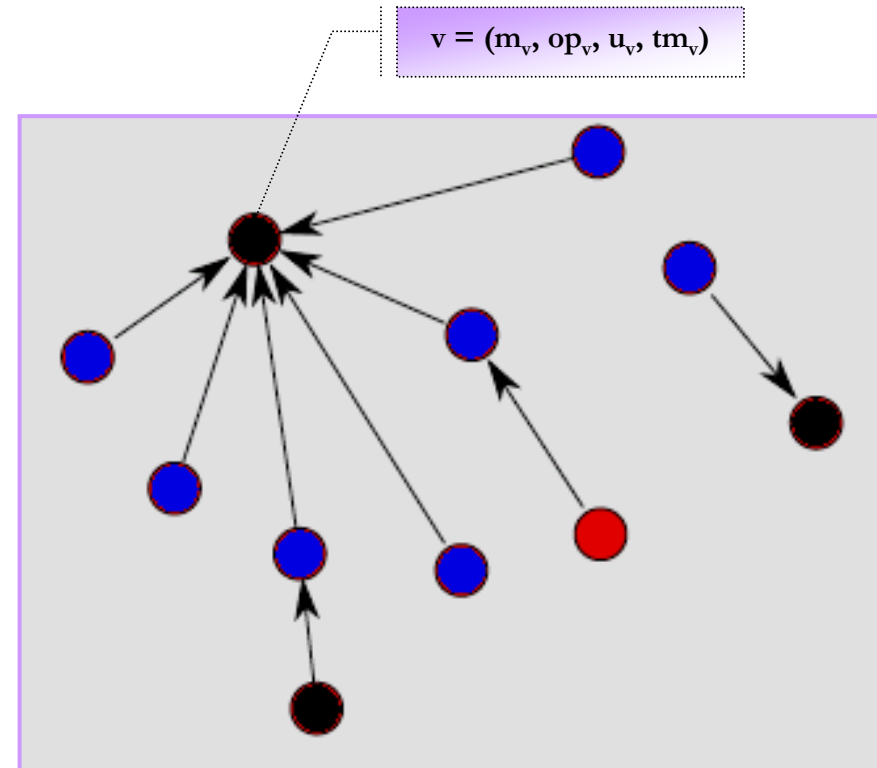
Mesures basées sur l'opinion

- **Opinion (Opinion Mining)**

- Statut d'un internaute
- **Statut vers un internaute**

$$avgOpToU sr(u) = \frac{\sum_i op_{v_i}}{\sum_j inDegree(v_j)}$$

- Réactions vers un message
 - Sommets d'une polarité
 - Opinion Moyenne
 - Variété de l'opinion



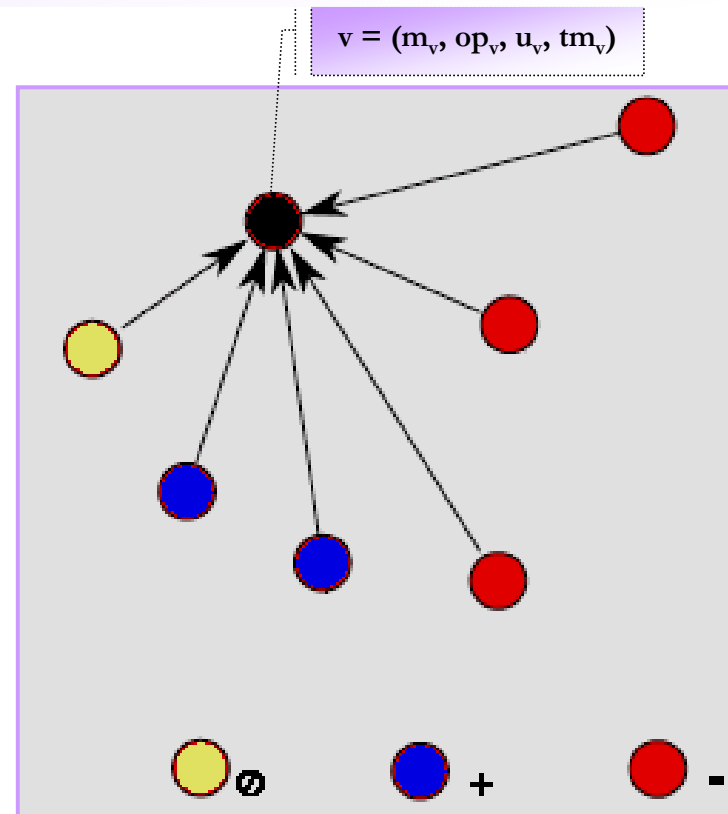
Mesures basées sur l'opinion

- **Opinion (Opinion Mining)**
 - Statut d'un internaute
 - Statut vers un internaute
 - **Réactions vers un message**
 - **Sommets d'une polarité**
 - **Opinion Moyenne**
 - **Variété de l'opinion**

$$\text{reply}(v, r) = |\{v' \in \text{inVertices}(v), \text{op}_{v'} = r\}|$$

$$\text{avgMsgOpinion}(v) = \frac{\sum_i \text{op}_{v'_i}}{\text{inDegree}(v)}$$

$$H(v) = - \sum_{r=n,o,p} \left(\frac{\text{reply}(v,r)}{\text{inDegree}(v)} \log \frac{\text{reply}(v,r)}{\text{inDegree}(v)} \right)$$



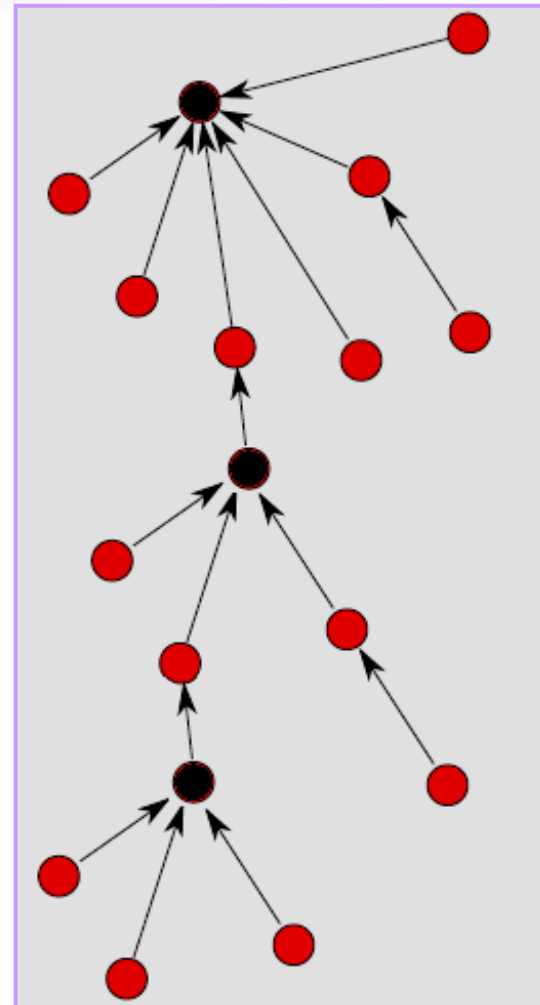
Mesures basées sur le thème

- **Thème (Topic Identification)**
 - **Evolution de l'opinion**
 - par utilisateur
 - dans le thème

$$opEvolution(u, T) = \frac{\sum_i op_{v_i}}{|msgs(u) \cap msgs(T)|}$$

$$opEvolution(u, T, tm_v, tm_{v'}) = |op_v - op_{v'}|$$

- Ancêtres



Mesures basées sur le thème

- **Thème (Topic Identification)**

- Evolution de l'opinion

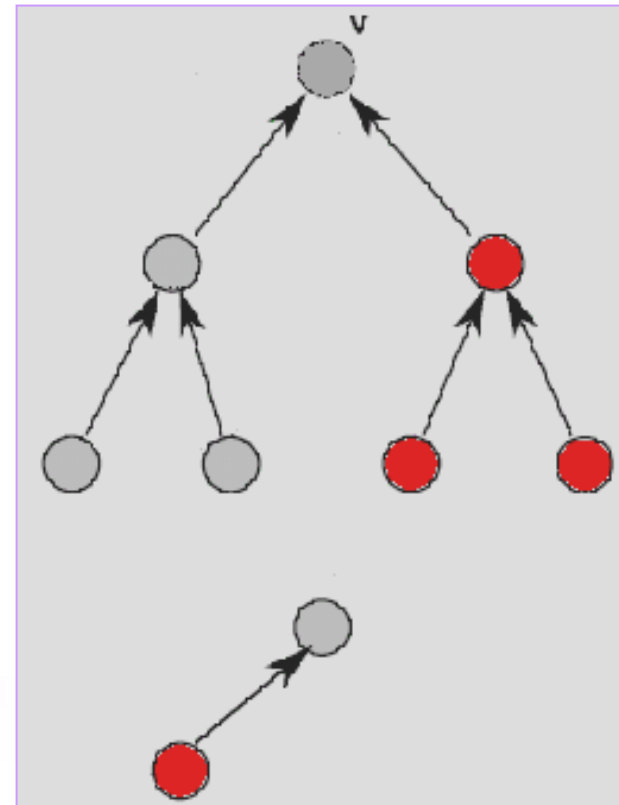
- par utilisateur
- dans le thème

$$opEvolution(u, T) = \frac{\sum_i op_{v_i}}{|msgs(u) \cap msgs(T)|}$$

$$opEvolution(u, T, tm_v, tm_{v'}) = |op_v - op_{v'}|$$

- **Ancêtres**

$$ancestors(v, T) = \{v' \in V : tm_{v'} < tm_v, topic(v') = topic(v)\}$$



Mesures basées sur le thème

- **Thème (Topic Identification)**

- Evolution de l'opinion

- par utilisateur

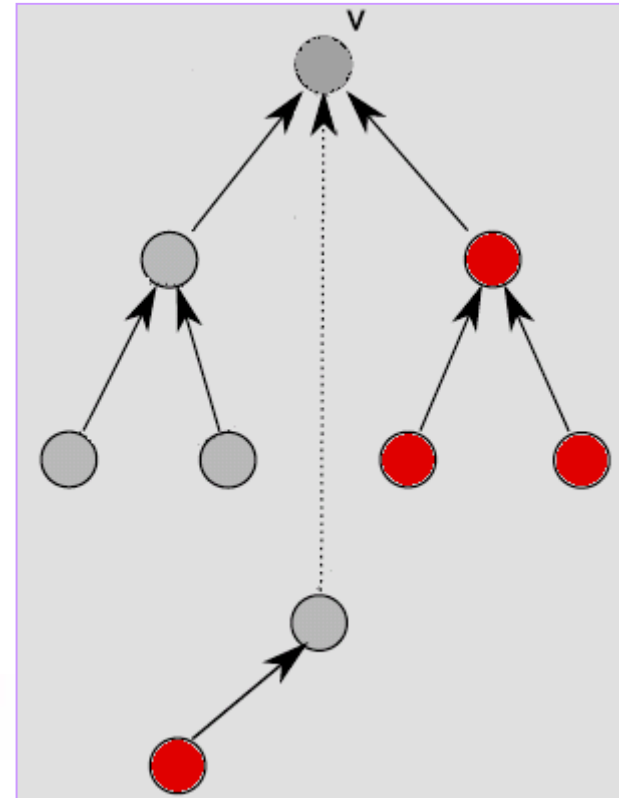
- dans le thème

$$opEvolution(u, T) = \frac{\sum_i op_{v_i}}{|msgs(u) \cap msgs(T)|}$$

$$opEvolution(u, T, tm_v, tm_{v'}) = |op_v - op_{v'}|$$

- **Ancêtres**

$$ancestors(v, T) = \{v' \in V : tm_{v'} < tm_v, topic(v') = topic(v)\}$$



Plan

- Méthodes existantes
- Notre proposition
- Utilisation du modèle
 - Mesures
 - **Recommandation**
- Prototype
- Conclusion – Perspectives

Systemes de recommandation

- Motivation : recommander des items (films, pages Web, nouveautés ...) aux internautes

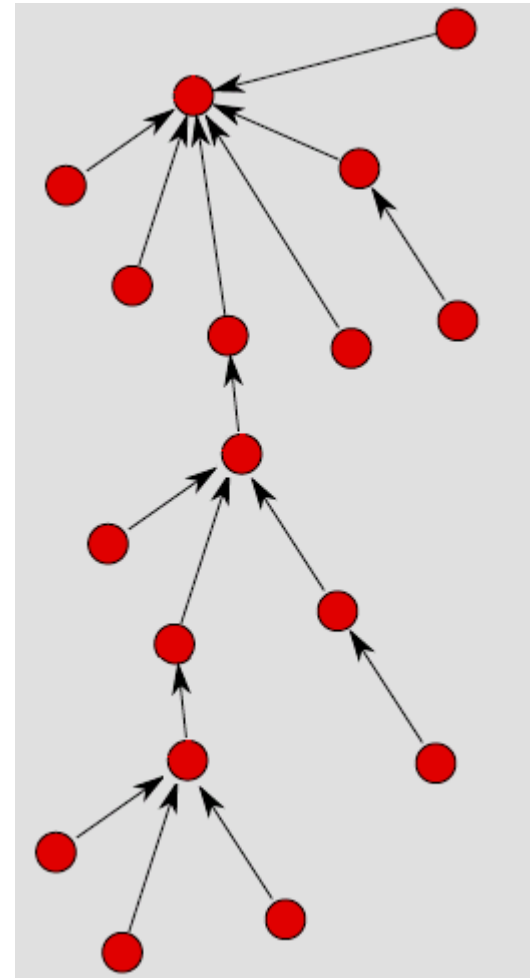
- Comment?
 - Votes par préférence (ex : étoiles)
 - 3 types
 - Basés sur le contenu (similarité entre items)
 - Collaboratifs (similarité entre internautes)
 - Hybrides

- Cas de « Cold-start »

Recommandation de messages intéressants

- **Nouvel internaute**
 - Par où commencer?
 - Où participer?

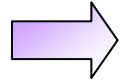
- **Cas de « Cold-start »**
 - Manque de profils



Messages Intéressants

- Interviewer 10 internautes qui visitent des discussions en ligne telles que les forums de manière quotidienne
 - ❑ **Opinion** : Un message est intéressant s'il contient des opinions.
 - ❑ **Taille** : Un message est intéressant s'il participe à un long fil de discussion.
 - ❑ **Réactions** : Un message intéressant a provoqué plusieurs réactions.
 - ❑ **Premier** : Le premier message d'un fil de discussion est intéressant.
 - ❑ **Chronologie** : Le message le plus récent est intéressant.

Messages Intéressants



■ Hypothèses

- Opinion
- Taille
- Réactions
- Premier
- Chronologie

■ Critères

- **OPINION** : l'opinion exprimée
- **ORDER** : nombre de sommets qui existent dans le fil de discussion où le message se trouve
- **POPULARITY** : la popularité du sommet
- **REPLY** : les réactions provoquées qui contiennent des opinions
- **ENTROPY** : la variété des opinions trouvées dans les réactions
- **ROOT** : s'il est le premier message du fil ou pas

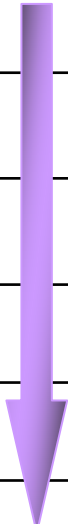
1ère expérience

- 8 forums du site <http://www.liberation.fr/forums/>
 - 1 147 messages – 636 réactions
- Objectif : les messages qui sont clés par rapport aux experts sont aussi classés comme clés par notre algorithme.
 - 2 experts, corrélation entre les humains et nos résultats
- Evaluation de notre algorithme
 - *précision* : éviter d'extraire des messages qui ne sont pas clés
$$precision = \frac{\text{correctly assigned key messages}}{\text{total key messages found by the system}}$$
 - *rappel* : combien des messages clés (notre méthode) sont considérés clés par les humains
$$recall = \frac{\text{correctly assigned key messages}}{\text{total key messages found by the user}}$$
 - F1-mesure $F1 = \frac{2*Recall*Precision}{Recall+Precision}$

Résultats 1ère expérience (par critère)

- Evaluation par critère
 - Précision basse : Variété du jeu de données
 - Probability Ranking Principle

CRITERE	F1 MOYEN
REPLY	0.30
POPULARITY	0.29
OPINION	0.28
ENTROPY	0.27
ROOT	0.26
ORDER	0.22



Résultats 1ère expérience (agrégation)

- Agrégation linéaire des critères, poids = 1

CRITERE	F1 MOYEN
REPLY	0.30
POPULARITY	0.29
OPINION	0.28
ENTROPY	0.27
ROOT	0.26
ORDER	0.22



F1 MOYEN AGREGATION
0.48

- Premier essai
- Jeu de données instable, experts

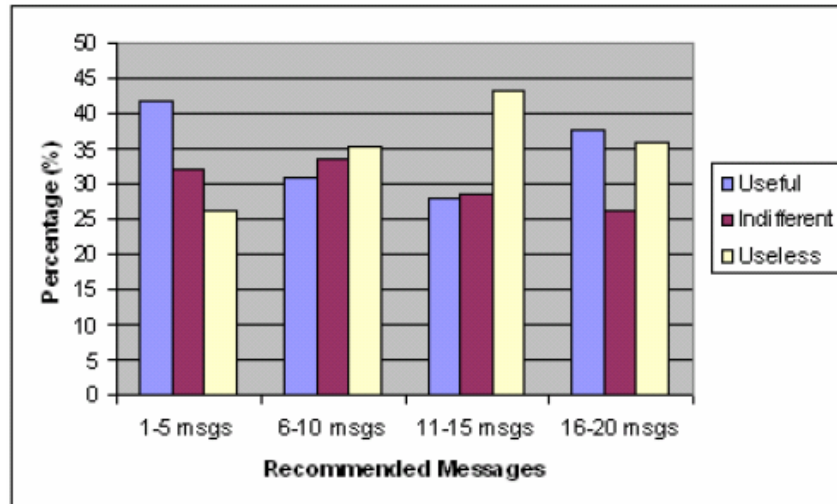
2ème expérience

- 6 experts
- forums anglais (<http://huffingtonpost.com>)
et français (<http://www.liberation.fr/>)
- **Objectif** : nous montrons les messages qui sont identifiés comme messages clés par des critères agrégés aux experts et ils indiquent s'ils sont d'accord ou pas.

Classement

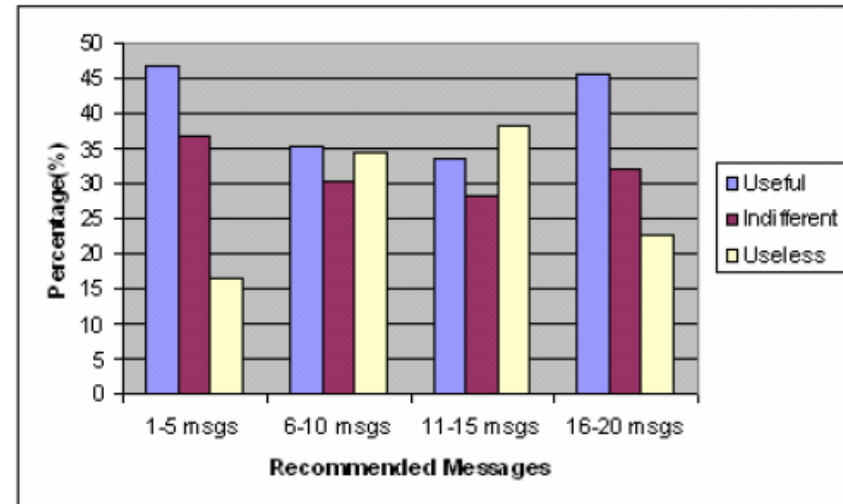
Utile	<ul style="list-style-type: none">■ Ce à quoi il répond■ Réactions qu'il a créées
Indifférent	<ul style="list-style-type: none">■ Antérieur/postérieur pas intéressants
Inutile	<ul style="list-style-type: none">■ Il n'aide pas■ Pas intéressant

Résultats 2ème expérience



Tous les messages

- 1-5, maximum de messages utiles, minimum d'inutiles
- > 10, beaucoup d'inutiles



Sans les messages courts

- Le nombre de messages inutiles diminue

Observations

- Personnalisation nécessaire
 - Profil de l'utilisateur (langue, préférences, croyances)
 - Style de forum (messages courts, opinion)

- Expériences avec des poids et des méthodes d'agrégation variées

Plan

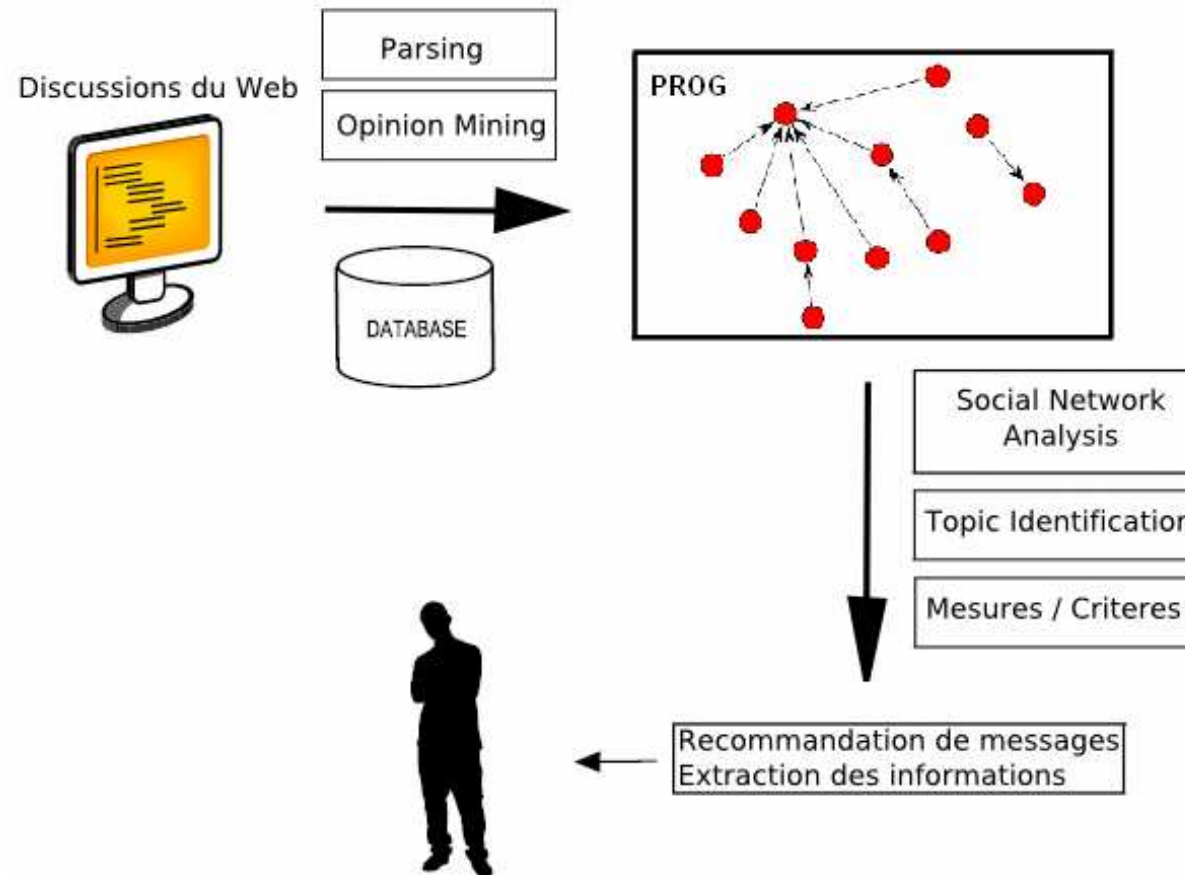
- Méthodes Existantes
- Notre Proposition
- Utilisation du modèle
 - Mesures
 - Recommandation
- **Prototype**
- Conclusion – Perspectives

Contexte

- Projet « Conversession »
 - Jeune entreprise soutenue par CREALYS (incubateur d'entreprises)

- Objectifs du projet :
 - Analyse de discussions en ligne
 - Synthétiser les tendances d'opinion

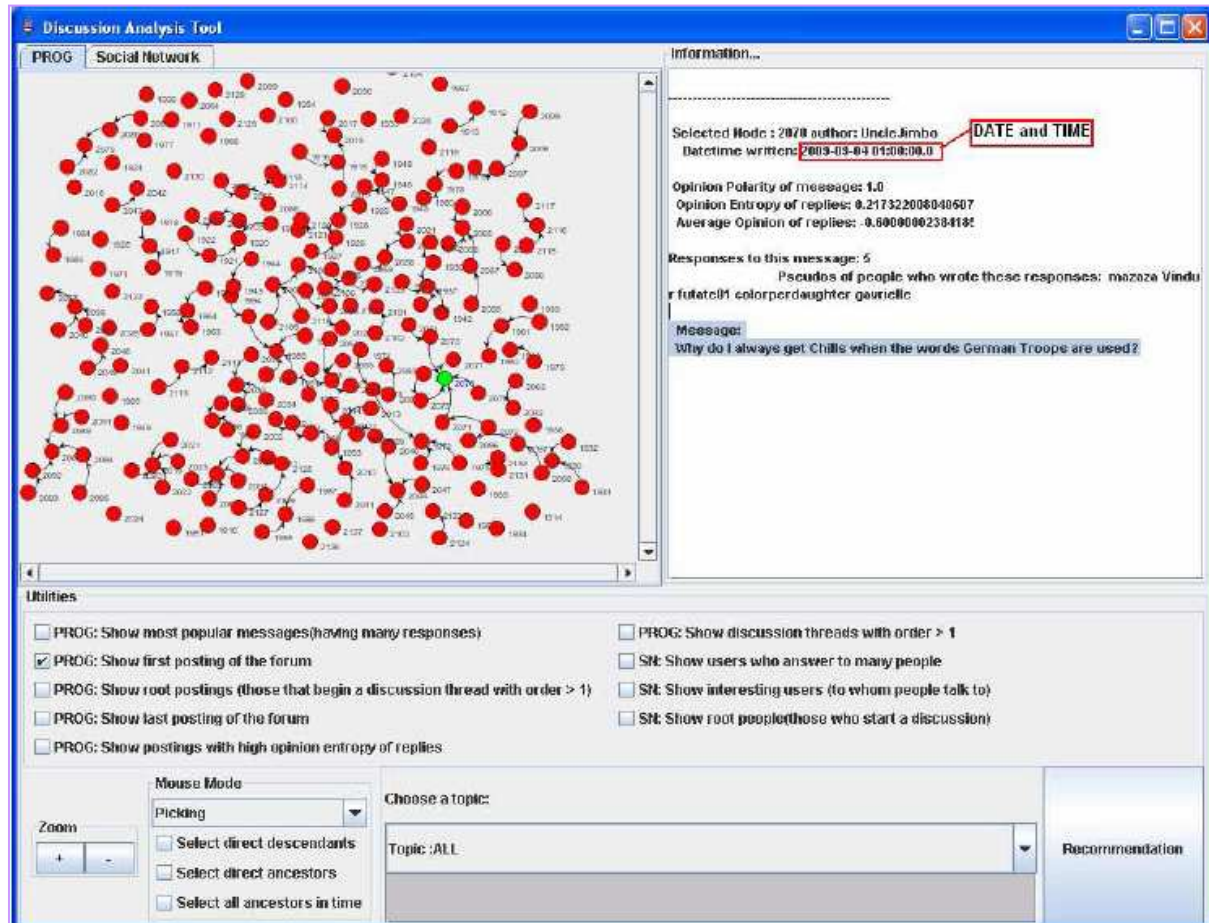
Prototype



Prototype

Discussion Analysis Tool

- Visualisation
- Navigation
- Mesures/Critères



Demonstration

DEMONSTRATION **Discussion Analysis Tool**

Titre : NATO jets bomb fuel tankers; Afghans say 70 killed

Messages : 228

Internauts : 118

Opinion Mining : SentiWordNet (A. Esuli and F. Sebastiani)

Conclusion - Contributions

■ Modèle et Mesures

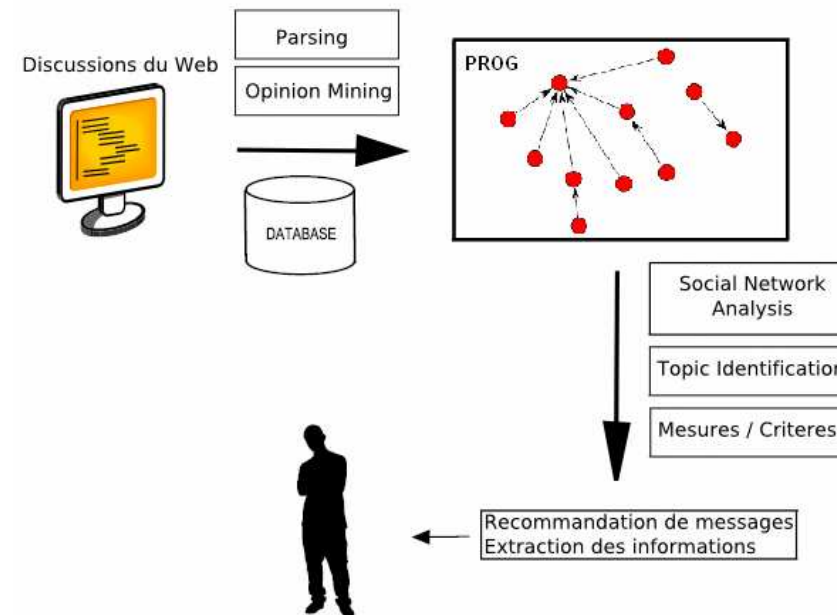
- Proposition d'un nouveau modèle
- Proposition de mesures pour l'extraction de connaissances

■ Recommandation

- Critères
- Evaluation

■ Prototype

- Multilingue
- Plugins pour la fouille de textes et d'opinions



Perspectives

■ **Modèle**

- ❑ **Combinaison** des modèles (réseaux sociaux et graphes de type PROG)
- ❑ **Evolution** de l'opinion
- ❑ **Mesures**

■ **Structure**

- ❑ Recherche automatique des **liens** entre les messages
- ❑ Un message peut répondre à **plusieurs** messages

■ **Recommandation**

- ❑ Personnalisation *explicite / implicite*
- ❑ Considérer un critère basé sur **PageRank**