Extraction d'opinions appliquée à des critères

Benjamin Duthil*, François Trousset*
Gérard Dray*, Pascal Poncelet**, Jacky Montmain*

*EMA-LGI2P, Parc Scientifique Georges Besse, 30035 Nîmes Cedex, France
prénom.nom@mines-ales.fr
**LIRMM UM2 - CNRS 5506, 161 Rue Ada, 34095 Montpellier Cedex 5, France
Pascal.Poncelet@lirmm.fr

Résumé. Les technologies de l'information et le succès des services associés (e.g., blogs, forums,...) ont ouvert la voie à un mode d'expression massive d'opinions sur les sujets les plus variés. Récemment, de nouvelles techniques de détection automatique d'opinions (*opinion mining*) ont fait leur apparition et via des analyses statistiques des avis exprimés, tendent à dégager une tendance globale des opinions exprimées par les internautes. Néanmoins une analyse plus fine de celle-ci montre que les arguments avancés par les internautes relèvent de critères de jugement distincts. Ici, un film sera décrié pour un scénario décousu, là il sera encensé pour une bande son époustouflante. Dans cet article, nous proposons, après avoir caractérisé automatiquement des critères dans un document, d'en extraire l'opinion relative. A partir d'un ensemble restreint de mots clés d'opinions, notre approche construit automatiquement une base d'apprentissage de documents issus du web et en déduit un lexique de mots ou d'expressions d'opinions spécifiques au domaine d'application. Des expériences menées sur des jeux de données réelles illustrent l'efficacité de l'approche.

1 Introduction

Avec le développement du Web, de plus en plus de documents textuels sont disponibles et de plus en plus d'outils permettent de rechercher de l'information pertinente. Connaître l'opinion des personnes sur un produit, rechercher et classer des documents, indexer de manière automatique des documents sont des problématiques d'actualité (e.g. Xu et al. (2011); Bai (2011); Morinaga et al. (2002)). Par exemple, dans le cas de l'opinion de cinéphiles, de nombreux outils sont disponibles pour connaitre l'avis général des spectateurs d'un film. Traditionnellement, pour extraire ces opinions, deux grandes approches existent : (i) celles basées sur un corpus d'apprentissage qui s'appuient généralement sur une analyse syntaxique et de co-occurrence des mots ; (ii) et celles utilisant un "dictionnaire" spécifique (e.g. *SenticNet* Cambria et al. (2010)) pour obtenir des orientations sémantiques d'un mot (Xu et al. (2011); Kamps et al. (2004)). Même si ces approches sont très efficaces, elles souffrent pour les premières de la nécessité de devoir constituer un corpus d'apprentissage. Dans des travaux précédents (Harb et al. (2008)), nous avons mis en avant le fait que l'expression d'opinions

pouvait tout à fait être différente en fonction du domaine considéré. Dans ce cadre, la définition d'un corpus est alors très difficile. Les approches utilisant un "dictionnaire" spécifique souffrent du même problème : un terme peut avoir une connotation positive dans un contexte et négative dans un autre. Enfin, traditionnellement ces approches se focalisent sur l'opinion générale exprimée dans un document et non pas sur l'opinion relative à chacun des critères spécifiques qui peuvent apparaître.
Récemment dans Duthil et al. (2011), nous avons proposé une approche d'extraction automatique de critères dans un document. Le principe général en est le suivant. A partir d'un ensemble de mots limité caractérisant un critère (e.g. réalisation, acteur), nous recherchons de manière automatique un corpus d'apprentissage et définissons à partir de celui-ci les différents mots pouvant être associés à un critère. Dans cet article nous montrons qu'un tel processus est tout à fait envisageable pour également extraire l'opinion exprimée dans les critères. Nous montrons également, que par rapport à Harb et al. (2008), qu'outre une classe positive ou négative, il est indispensable de définir une nouvelle classe (appelée anti-classe) afin de mieux caractériser les opinions.

2 Acquisition du corpus d'apprentissage

L'objectif de cette phase est d'acquérir automatiquement des textes porteurs d'opinions pour construire un corpus d'apprentissage nécessaire à l'apprentissage et la classification des mots exprimant une opinion positive, négative. De manière à extraire un ensemble de documents significatifs, l'ensemble des documents est obtenu à partir du Web. Pour assurer la présence d'opinions positives et négatives dans les textes constituant le corpus d'apprentissage, nous recherchons des textes contenant au moins un mot porteurs d'opinion. Nous considérons deux ensembles P et N de mots porteurs d'opinions, respectivement positives et négatives, appelés *mots germes* utilisés dans la littérature (Turney (2002)) : $P = \{$*good*, *nice*, *excellent*, *positive*, *fortunate*, *correct*, *superior*$\}$ et $N = \{$*bad*, *nasty*, *poor*, *negative*, *unfortunate*, *wrong*, *inferior*$\}$; Pour chaque mot germe de l'ensemble P (resp. N) et pour une thématique donnée $\mathcal{T}$, nous collectons automatiquement K documents d'opinion relatifs à $\mathcal{T}$ (par exemple *movie*) et qui contiennent l'opinion exprimée par le mot germe $g \in \{P \cup N\}$, soit 14 corpus S_g de K documents.

3 Extraction des descripteurs d'opinions

À partir des corpus collectés, nous cherchons des corrélations entre les mots germes et d'autres adjectifs et expressions (i.e. la concaténation des adverbes précédents un adjectif), dans les documents. Nous appellerons par la suite les mots appris décrivant une opinion : "*descripteurs*". Ces descripteurs expriment une opinion positive ou négative. Le but de l'apprentissage est d'enrichir les ensembles de mots germes par des descripteurs qui expriment une opinion proche de celle des mots germes. Nous considérons que l'opinion exprimée par des descripteurs proches d'un mot germe sont similaires à celle exprimée par ce dernier (ou du moins de même polarité (positif/négatif)). Ainsi, plus un descripteur est corrélé, i.e, qu'il apparaît proche d'un mot germe, plus il est vraisemblable qu'il ait la même polarité que le mot germe auquel il est associé. De la même manière, nous considérons les descripteurs "éloignés"

(loin d'un mot germe) comme non pertinents pour le mot germe. À chaque mot germe est associée une liste de mots qui lui sont fortement corrélés, on appelle cet ensemble de mots la *classe* C du mot germe. De même, si l'on considère une liste de mots qui ne sont pas corrélés, on appelle cet ensemble de mots l'*anti-classe* AC du germe. Cette technique de discrimination permet de supprimer les descripteurs non pertinents, en étudiant la corrélation de chacun d'eux dans la classe et dans l'anti-classe (fréquence). Si un descripteur est plus fréquent dans la classe que dans l'anti-classe, il sera considéré comme pertinent et de même polarité que les mots de la classe C. À l'inverse un descripteur plus fréquent dans l'anti-classe que dans la classe sera considéré comme non pertinent avec une polarité opposée. Si le descripteur est corrélé de la même manière à la classe et à l'anti-classe, il est alors considéré comme non pertinent.

Pour assurer la notion de proximité précédemment décrite dans le calcul de corrélation, nous introduisons la notion de fenêtre F de taille sz centrée sur un mot germe g d'un document t appartenant à S_g (les textes rattachés au germe g) par : $F(g, sz, t) = \{m \in t / d^t_{JJ}(g, m) \leq sz\}$ où $d^t_{JJ}(g, m)$ est la distance correspondant au nombre d'adjectifs (JJ) séparant le mot m de g.

Pour calculer la corrélation entre les descripteurs appartenant à l'ensemble des classes de chacun des germes appartenant à P : C_P (resp. appartenant à C_N) par rapport à un mot germe g nous utilisons une fenêtre centrée sur g. Nous calculons leur fréquence d'apparition $X(M)$ dans chacune des fenêtres des textes appartenant à $S_P = \cup(S_g / g \in P)$ (resp. $S_N = \cup(S_g / g \in N)$ où S_P est l'ensemble des textes contenant les mots germes positifs. On appelle $\mathcal{O}(g, t)$ les occurrences du germe dans le texte t et $\mathcal{O}(m, F(g, sz, t))$ le nombre d'occurrences de m dans la fenêtre $F(g, sz, t)$ d'un texte t. $X(M)$ est alors défini par :

$$X(M) = \sum_{g \in \mathcal{P}} \sum_{t \in S_g} \sum_{\gamma \in \mathcal{O}(g,t)} \mathcal{O}(M, F(\gamma, sz, t)) \tag{1}$$

Ce qui correspond à cumuler les fréquences d'un descripteur présent dans toutes les fenêtres présentes dans tous les textes appartenant à S_P (resp. S_N), i.e tous les textes associés aux germes appartenant à P (resp. N). De la même manière, nous calculons la corrélation des descripteurs appartenant à AC_P (resp. AC_N) par rapport à un mot germe g (fréquence d'apparition à l'extérieur des fenêtres $\overline{F'}$ centrées sur g), $\overline{X}$ est alors défini par :

$$\overline{X}(M) = \sum_{g \in \mathcal{P}} \sum_{t \in S_g} \sum_{\gamma \in \mathcal{O}(g,t)} \mathcal{O}(M, \overline{F'}(\gamma, sz, t)) \tag{2}$$

À partir des fréquences $X(M)$ et $\overline{X}(M)$ nous pouvons déterminer l'appartenance d'un descripteur à C et AC en calculant son score tel que $Sc(M) = X - \overline{X}$.

Nous en déduisons la polarité des descripteurs : si $Sc(M) < 0$ alors $M \notin C$ et si $Sc(M) > 0$ alors $M \in C$; $C \in \{C_P, C_N\}$.

4 Utilisation des descripteurs d'opinions

À partir des descripteurs obtenus précédemment nous cherchons à déterminer l'opinion qui se dégage d'un texte, ou d'une portion de texte, en rapport avec la thématique $\mathcal{T}$.
Pour un texte t, nous utilisons une fenêtre glissante de taille sz successivement centrée sur chaque occurrence d'un adjectif dans le texte t. Un score est calculé pour chacune des fenêtres

f de la manière suivante : $Score(f) = \sum_{M \in f} Sc(M)$. La polarité de t est déterminée par le signe de $Sc(t)$ défini par $Sc(t) = \sum_{f \in t} Score(f)$. Si $Sc(t) < 0$, t est négatif, si $Sc(t) > 0$, t est positif.

5 Expérimentations

Pour valider notre approche de détection d'opinions nous avons choisi d'utiliser le corpus de critiques cinématographiques proposé par Pang et al. (2002) et comme thématique $\mathcal{T}$ le cinéma. Chacun des textes est étiqueté comme positif ou négatif. Pour valider l'approche dans un contexte utilisant des critères, nous avons extrait les parties de texte relatives aux critères "acteur" et "scénario" en utilisant la technique appelée *Synopsis* proposée par Duthil et al. (2011), puis nous détectons l'opinion relative à chacun d'eux.
Pour comparer nos résultats, nous avons choisi un outil de détection d'opinions de la littérature : *SenticNet* (Cambria et al. (2010)), qui propose une collection de concepts polarisés (positifs/négatifs) constituant un réseau sémantique. Les indicateurs classiques de *précision*, *rappel* et *FScore* sont utilisés pour évaluer les performances des systèmes.

5.1 Validation de l'approche en classification de textes

Nous proposons ici d'évaluer l'approche dans un contexte de classification de textes.

	Notre approche		**SenticNet**	
	positif	***négatif***	***positif***	***négatif***
FScore	**0,73**	**0,69**	0,68	0,68
précision	**0,68**	**0,75**	0,54	0,74
rappel	0,79	**0,63**	0,91	0,25

TAB. 1 – *Résultats sur le corpus de Pang et al. (2002) en classification de textes*

.

Le tableau 1 met en évidence l'efficacité de l'approche dans un contexte de classification de documents. On remarque que le $FScore$ obtenu par notre approche est meilleur que celui obtenu par *SenticNet*. On observe une petite faiblesse de notre approche dans la détection de textes positifs, mais qui est largement compensée par la détection de textes négatifs (le *rappel* (0.63) est plus de deux fois plus important que celui de *SenticNet* (0.25)). Ces résultats mettent en évidence qu'un vocabulaire d'opinions non spécifique à une thématique pourrait suffire pour une évaluation globale d'un document. Nous pouvons constater que notre approche obtient des résultats similaires par rapport à une détection universelle pour la classification de textes (positifs/négatifs) et que le lexique appris est pertinent. Le seul apprentissage des adjectifs et expressions est donc suffisant pour réaliser cette tâche. Par ailleurs, nous pouvons souligner que notre méthode est entièrement automatique, et nécessite seulement une expertise minimale contrairement à *SenticNet* qui est entièrement supervisée.

5.2 Validation de l'approche sur deux critères, ici les critères "*acteur*" et "*scénario*"

Nous proposons ici d'évaluer l'extraction d'opinions par rapport à des critères.

	Critère *Acteur*				**Critère *Scénario***			
	Notre approche		**SenticNet**		**Notre approche**		**SenticNet**	
	positif	***négatif***	***positif***	***négatif***	***positif***	***négatif***	***positif***	***négatif***
FScore	**0,92**	**0,92**	0,70	0,70	**0,83**	**0,80**	0,69	0,69
précision	**0,90**	**0,95**	0,60	0,74	**0,76**	**0,90**	0,55	0,74
rappel	**0,95**	**0,90**	0,87	0,38	**0,92**	**0,71**	0,91	0,26

TAB. 2 – *Résultats en classification de textes sur les critères "acteur" et "scénario"*

.

Le tableau 2 montre que l'opinion exprimée par les internautes dans des critiques cinématographiques dépend de critères. Ici, on remarque que le critère "acteur" correspondrait mieux au point de vue des personnes ayant rédigé les critiques que le critère "scénario" pour lequel les résultats de la classification sont moindres. Dans une approche multicritère de l'évaluation d'opinions, on pourrait par exemple en déduire que le critère acteur pourrait avoir un poids important pour la décision d'un avis sur un film. Les résultats du tableau 1 correspondent à un score d'opinion de synthèse et que, au vu des résultats, nous mettons en évidence que la méthode d'agrégation qu'a l'esprit humain ne se limite pas à une simple moyenne arithmétique, mais plutôt qu'il agit selon un système de préférences complexes. Nous pouvons remarquer que dans un contexte d'extraction d'opinions relatives à des critères, *SenticNet* est beaucoup moins efficace que notre méthode pour plusieurs raisons. Le fait d'introduire des critères a mis en évidence que le vocabulaire d'opinions est spécifique à la thématique, et que la combinaison de notre approche de détection d'opinions à un processus d'extraction thématique non supervisé (Duthil et al. (2011)), requiert une expertise minimale contrairement à un processus supervisé comme *SenticNet*.

6 Conclusion

Dans cet article nous avons proposé une nouvelle approche permettant d'extraire automatiquement, avec une expertise minimale, les opinions présentes dans des textes par rapport à des critères. D'autre part, notre approche permet de construire automatiquement un lexique de descripteurs d'opinions relatifs à une thématique. Nous avons illustré la complexité qu'a l'esprit humain à construire une opinion, et qu'elle repose sur des critères propres à chaque individu, ce qui rend la tâche d'automatisation cognitive difficile. D'autre part, nous avons proposé une approche permettant la construction automatique d'un corpus d'apprentissage, ainsi qu'une technique efficace de classification des mots d'opinions. Les perspectives associées à ce travail sont nombreuses. Tout d'abord, nous souhaitons étendre nos expérimentations à plusieurs thématiques pour montrer la réelle dépendance entre les descripteurs et la thématique dans laquelle ils sont utilisés. Nous souhaitons intégrer à notre apprentissage les verbes et les noms qui sont aussi porteurs d'opinion, et qui permettent souvent de nuancer l'opinion exprimée.

Extraction d'opinions.

Références

Bai, X. (2011). Predicting consumer sentiments from online text. *Decis. Support Syst. 50*, 732–742.

Cambria, E., R. Speer, C. Havasi, et A. Hussain (2010). Senticnet : A publicly available semantic resource for opinion mining. *Artificial Intelligence*, 14–18.

Duthil, B., F. Trousset, M. Roche, G. Dray, M. Plantié, J. Montmain, et P. Poncelet (2011). Towards an automatic characterization of criteria. In *DEXA (1),LNCS*, pp. 457–465.

Harb, A., M. Plantié, G. Dray, M. Roche, F. Trousset, et P. Poncelet (2008). Web opinion mining : how to extract opinions from blogs ? *International Conference on Soft Computing as Transdisciplinary Science and Technology*.

Kamps, J., M. Marx, R. J. Mokken, et M. de Rijke (2004). Using wordnet to measure semantic orientations of adjectives. In *In Proceedings of LREC-04, 4th international conference on language resources and evaluation, Lisbon, PT*, Volume 4, pp. 1115–1118.

Morinaga, S., K. Yamanishi, K. Tateishi, et T. Fukushima (2002). Mining product reputations on the web. In *In ACM SIGKDD 2002*, pp. 341–349. ACM Press.

Pang, B., L. Lee, et S. Vaithyanathan (2002). Thumbs up ? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Turney, P. (2002). Thumbs up or thumbs down ? semantic orientation applied to unsupervised classification of reviews. *In Proceedings of 40th Meeting of the Association for Computational Linguistics*, 417–424.

Xu, K., S. S. Liao, J. Li, et Y. Song (2011). Mining comparative opinions from customer reviews for Competitive Intelligence. *Decision Support Systems 50*(4), 743–754.

Summary

The development of new services (e.g. blogs, forums, etc) provides facilities to express opinion on different topics. Recently new techniques known as *opinion mining* have emerged and use statistical analysis tools to extract an overall trend of the opinions expressed by users on different topics. Nevertheless with such a trend they are not able to focus on a more detailed analysis on criteria. For example, a film can be classified as positive, while the general opinion expressed on the music criteria is negative. Our objective in this paper is to automatically extract the opinion expressed in fragments of texts. From a small set of opinion keywords, we show how to automatically construct a learning base and then extract significant ways of expressing opinions. Experiments conducted on real datasets illustrate the effectiveness of our proposal.