Data Protection for Recommendation Engines: Obstacle or Opportunity?

SPIDEO

SIMPLY RELEVANT

Summary

In the new media industry, users' personal data is valuable, particularly when it comes to content recommendation. Users and regulators value privacy according to basic principles. While data protection is seen as an obstacle by some, Spideo believes it's actually a springboard. However, from a data privacy standpoint, not all recommendation technologies were created equal. By moving the focus away from only scrutinizing users and closer to a deep understanding of content, content-based filtering methods enhanced with semantics improve adoption and cement loyalty thanks to added transparency and increased trust from the users.

Introduction

Whether it's called Big, Smart or Personal, data is at the core of the new media industry. Both medium and outcome of new ways of distributing and consuming content, the intrinsic value of data generated by users is high, but so are concerns over use of personal data and privacy breaches. Data is truly a word on everyone's lips and as a result the stakes are particularly high in any conversation about the topic.

Processing data represents sizeable advantages and economic opportunities for operators since it gives them clear and accountable intelligence that they will use in turn to invest in content, and for marketing purposes.

As explained to The New York Times by an anonymous Amazon executive in 2013: "It is clear that having a very molecular understanding of user data is going to have a big impact on how things happen in television"[1]. The value of media companies is now not only found in viewing figures, but also in the intelligence that they possess about their users and subscribers.

However, data also carries with it the risk of breaching people's privacy and the general public is often concerned about unwarranted use of its data by businesses.

Symantec published a study in 2015 showing that 66% of European consumers were calling for measures to improve personal data protection[2].

In the new media economy, if content providers want to invest sustainably in their brand, they will have to face the public on the issue of privacy. Well aware of this public debate, the EU adopted last April a new General Data Protection Regulation[3] that will come into full effect in May 2018.

As a Recommendation and Analytics company, no one more than us knows how personal data can be so truly and deeply reflective of individuals' tastes and personality. Therefore, it is important for us to know our end-users well but without being perceived as intrusive, in order to build a trusting relationship with them. Actually, we chose not to see data protection as an obstacle or a wall, but as a guiding principle advantageous for all parties to adhere to, as will be made apparent in this White Paper.

With a perspective informed by data protection principles, one thing quickly becomes apparent: when it comes to respecting privacy, not every recommendation engine was created equal.

▶ **Which technologies currently provide personalized recommendations that are non-intrusive and ultimately user friendly?**

▶ **How can compliance with data protection rules reassure users and benefit service providers?**

▶ **How can Recommendation & Analytics companies approach Data Protection not as a hindrance but as an advantage?**

[1] CARR (D.), "Giving viewers what they want", The New York Times, Feb. 2013. http://www.nytimes.com/2013/02/25/business/media/for-house-of-cards-using-big-data-to-guarantee-its-popularity.html?_r=0

[2] SYMANTEC, "State of privacy", 2015. https://www.symantec.com/content/en/us/about/presskits/b-state-of-privacy-report-2015.pdf

[3] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

PERSONAL DATA FOR DUMMIES

The EU's definitions

According to Article 4 of the EU's General Data Protection Regulation, personal data is any information relating to an identified or identifiable natural person (the "data subject") who can be "identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person";

7 Principles defined by the GDPR:

▶ **Lawfulness, fairness and transparency**

Personal data must be processed lawfully, fairly, and in a transparent manner in relation to the data subject.

▶ **Purpose limitation**

Personal data must be collected for specified, explicit and legitimate purposes and not further processed in a way incompatible with those purposes.

▶ **Data minimisation**

Personal data must be adequate, relevant and limited to those which are necessary in relation to the purposes for which they are processed.

▶ **Accuracy**

Personal data must be accurate and, where necessary, kept up to date.

▶ **Storage limitation**

Personal data must be kept in a form which permits identification of data subjects for no longer than is necessary.

▶ **Integrity and confidentiality**

Personal data must be processed in a manner that ensures appropriate security of the personal data.

▶ **Accountability**

The controller shall be responsible for and be able to demonstrate compliance with these principles.

1. What are the existing technologies which provide non-intrusive personalized recommendations?

Recommendation algorithms can be classified in three main categories.

Social graphs use global mapping of users based on the largest number of defined relationships possible and how they're related to each other. Content recommendation engines based on social graphs commonly source user's data from third party social networks, especially Facebook's valuable Open Graph API which maps user relationships of 1.65 billion monthly users.

Collaborative filtering methods are based on collecting and analyzing a large amount of information based on users' behaviors, activities or preferences and predicting what users will like based on their similarity to other users. It is based on statistical models and does not require an "understanding" of the recommended item itself.

Instead of focusing exclusively on users, **content-based filtering** methods are based on a specific knowledge and description of the recommended items and in our case, content. This knowledge goes from traditional tagging methods (flat and unidirectional) to complex ontologies (weighted and multi-directional semantic networks).

1.1 How do recommendation technologies fit with data protection principles?

Drawing inspiration from the legal concepts developed by the EU regulation and taking them into account from a user perspective, we observe that three principles make a real difference for quality of service:

▶ **Relevance:** Personal data must be adequate, relevant and not excessive

▶ **Necessity:** Only personal data necessary for the purposes of recommendation should be collected and processed.

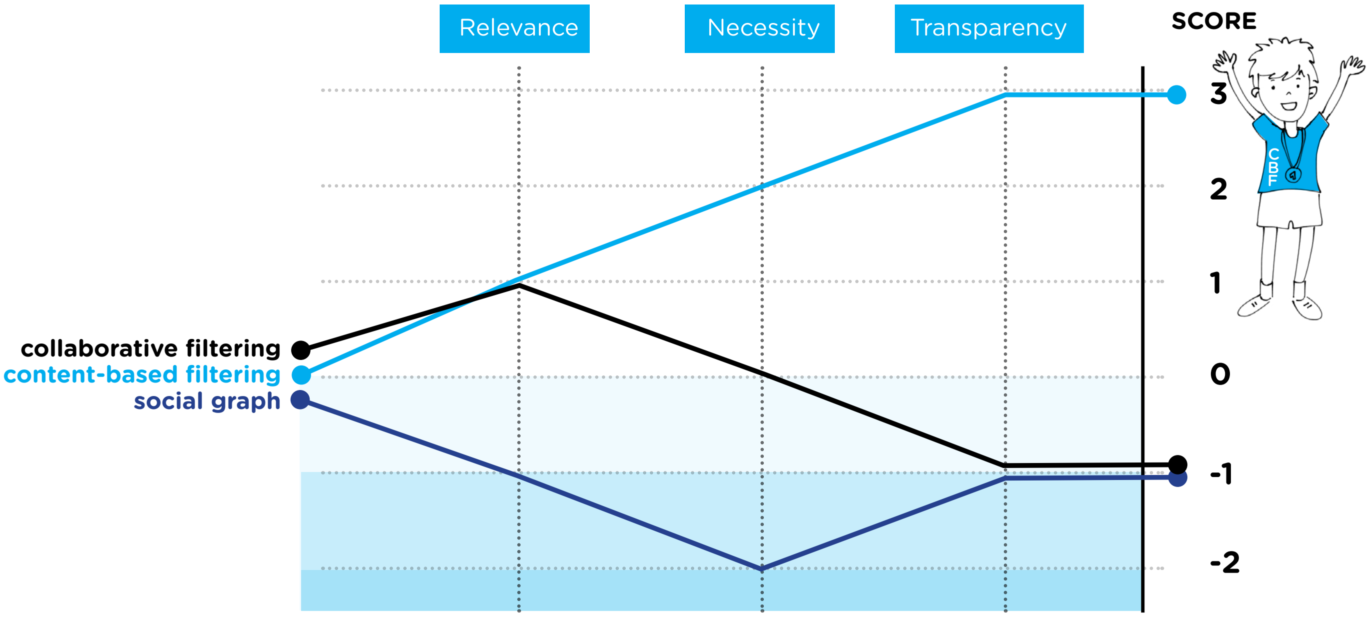
▶ **Transparency:** The collection of personal data requires personal control and feedback to users.

In practice, social graphs and collaborative filtering methods are potentially in contradiction with more than one of these three principles.

Recommendation engines that are based both on social graphs and collaborative filtering methods have a user-related data approach. They compensate for their lack of relevant information about content (content-related data) through an overabundance of information about users. This often goes way beyond necessity and proportionality requirements and thus strongly undermines their capacity to be in compliance with data protection principles.

Indeed, technologies that feed their algorithms mainly with social graphs use all sorts of information, much of which is completely disconnected from the video watching experience. While there is value in knowing basic demographic criteria such as age, gender and location, other pieces of information are scrutinized such as a user's friends, the communities they engage with and apps used, to name but a few. One must wonder to which extent is having access to this breadth of information is relevant when it comes to recommending video content.

Statistical methods like collaborative filtering depend on the law of large numbers: they need to be deployed to vast numbers of users in order to work efficiently. Consequently, this creates strong incentives to prevent users from opting out of tracking mechanisms by building indirect obstacles beyond their control. Simultaneously, since collaborative filtering algorithms are not interested in the intrinsic characteristics of what they actually recommend, they can only provide little feedback to users, often of the "Other customers also bought" type, which could go against transparency principles.

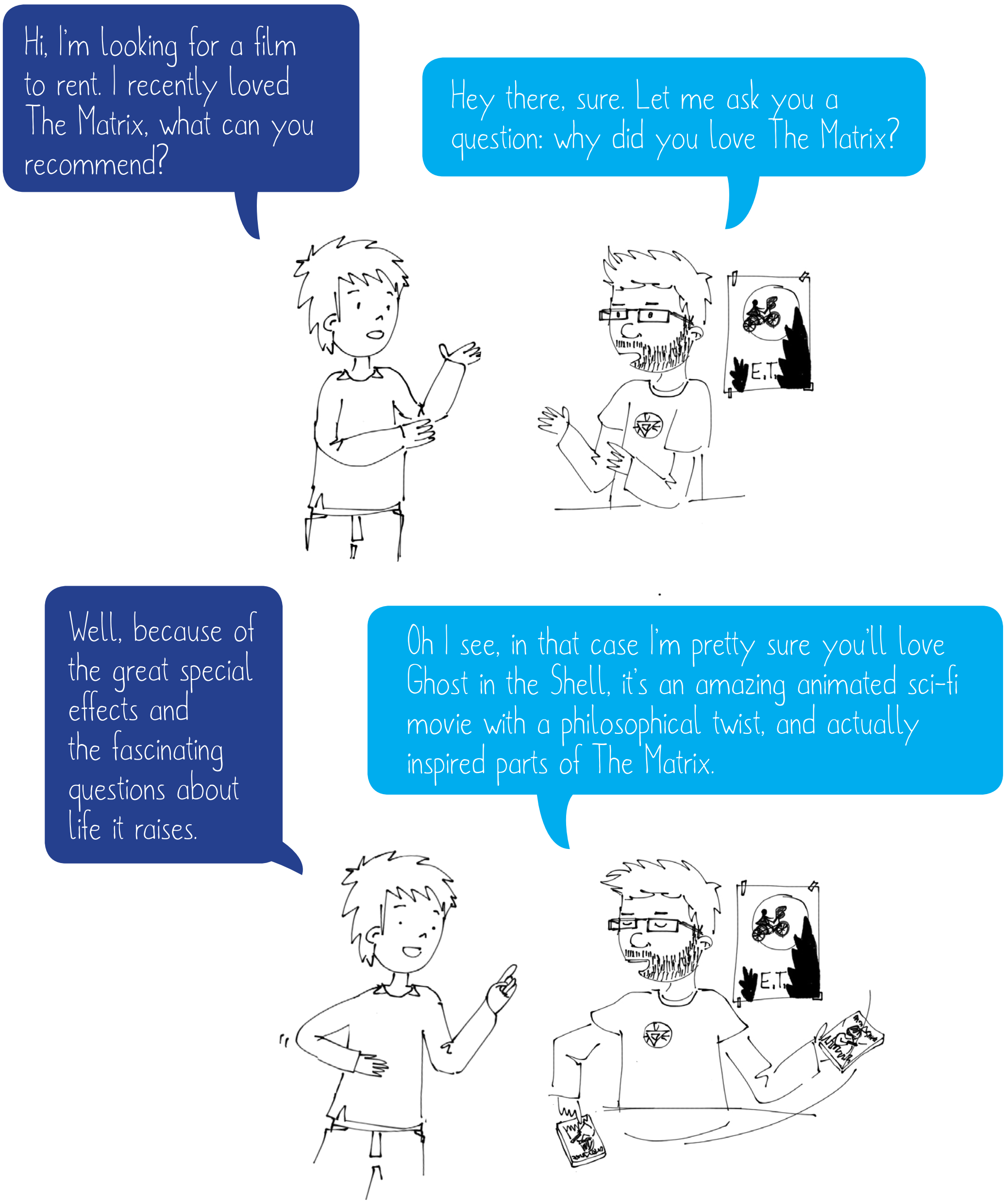


1.2 How about focusing less on users and more on content?

With their inherent focus on knowing as much as possible on content instead of endlessly scrutinizing users, content-based filtering methods are respectful of data protection principles. In order to understand why, let's momentarily go back to the world which existed before big data was a buzzword.

Back in the 20th century, going to the video store often was the most convenient option when we wanted to watch a movie. Many of us are familiar with feeling overwhelmed by the sheer possibilities we were faced with as we walked inside the store. Between wandering aimlessly through the aisles for hours or interacting with knowledgeable video clerks for much needed advice, the second option was always more efficient.

Conversations usually went a little something like this:


Hi, I'm looking for a film to rent. I recently loved The Matrix, what can you recommend?
Hey there, sure. Let me ask you a question: why did you love The Matrix?
E.T.
Well, because of the great special effects and the fascinating questions about life it raises.
Oh I see, in that case I'm pretty sure you'll love Ghost in the Shell, it's an amazing animated sci-fi movie with a philosophical twist, and actually inspired parts of The Matrix.
E.T.

Very little personal information is revealed in this short conversation. The video rental clerk here knows almost nothing about that person, their age, what part of town they live in, who their friends are or where they usually go for dinner. Technically speaking, that person's anonymity is conserved.

The video store clerk is a bona fide movie buff, so not only does he have deep knowledge of each title available in the store, he also perfectly understands how they relate to each other. The combination of both is insightful enough to provide enough information in order to make the best recommendation possible. Transparent feedback provided when explaining why a particular film is recommended shows the customer that the clerk "gets it" and paves the way for a great customer relationship in the future.

While some other recommendation services seek out as much information as possible about their users (more than often handling data that is of debatable relevance for video recommendation) we actually don't need to know too much about users beyond their actual tastes and preferences.

2. Why is it in everyone's best interest to comply with data protection principles?

2.1 How seriously should financial sanctions be taken?

Up until now, businesses who infringe on data protection rules do not seem to be worried by the threat of financial sanctions. This is especially true in Europe as these sanctions are capped to amounts which don't always encourage taking the issue seriously (i.e. France's article 47 de la loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés and the 150 000 EUR fine in case of failure to comply or Germany's 50,000 to 300,000 fine as part of (§ 43 III BDSG). Application criteria of these sanctions is also limited in the sense that European regulators cannot penalize companies established outside of the EU's territory that don't rely on data processing methods within that territory[4].

Public opinion's growing concern on matters of personal data is nonetheless making lines shift. Thus, the EU's General Data Protection Regulation calls for tougher financial sanctions:

▶ Some infringements of the Regulation provisions will be subject to administrative fines of up to 10,000,000€ or of up to 2% of the total worldwide annual turnover;

▶ The most prejudicial infringements of the Regulation provisions will be subject to administrative fines of up to 20,000,000€ or of up to 4% of the total worldwide annual turnover"[5].

In addition, it also intends to broaden sanction applicability to any data controller whose processing activities are directed to EU residents[6].

Undoubtedly, while the issue of financial risk through sanctions could currently almost be dismissed as incidental, it is about to become a very serious matter. The threat of considerably tougher sanctions will force economic parties involved in the use of personal data to be much more vigilant in the future.

However, the core issue raised by personal data protection is located elsewhere. It's about reputation and brand perception.

With this perspective in mind, regulators have become more and more interested in communication sanctions, rather than financial sanctions.

80% of Europeans do not trust retailers with their personal data[7].

So the big question is: how to go about gaining and maintaining consumer trust? Failing to provide answer threatens the very raison d'être of services that rely on personal data.

[4] Article 4 of Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data.

[5] Article 83 of the General Data Protection Regulation

[6] Article 3 of the General Data Protection Regulation

[7] SYMANTEC, "State of privacy", 2015. https://www.symantec.com/content/en/us/about/presskits/b-state-of- privacy-report-2015.pdf

2.2 What about user sensibility?

> “*Instead of drowning personal data protection within End-User License Agreements which are too often vague and too complicated to properly understand, it needs to become the very foundation of a trusting relationship with users. Advanced features require access to a person’s location? To their movements? To their travelling speed? To the sensors on their phones and their connected objects? Tomorrow, to their emotions and moods? Why not... but what for?”[8].*
>
> Isabelle Falque-Pierrotin, Chair of the CNIL, the French data protection authority.

The CNIL recently highlighted user experience as being at the heart of the topic rather than approaching it through a purely regulatory perspective. This is a particularly interesting side step because it moves the focus onto the industry’s own arena: service quality and user adoption.

Commercial successes and use successes of personalized video services be built against the best interests of consumers.

They must go through full and total acquiescence of users and a clear understanding of the value of services offered.

A recent example from the music sector is revealing of user’s sensitivity relating to their personal data and its uses. In August of 2015, Spotify subscribers diligently rallied to express their discontent following modifications made by the company to its personal data treatment policy. The company had just informed its users that the platform would take the liberty of collecting information recorded on their mobile devices such as their “contacts, pictures and multimedia files”. This decision unleashed a chain reaction of criticism and many have virulently pointed out their inability to understand how such information could ever be useful to a music service[9].

Spotify’s reputation has considerably suffered from this episode that shows how much users value relevance, proportionality and transparency (the three basic data protection principles) when it comes to their data. Tangible and harmful consequences of this sort of negligence are engagement loss and unsubscriptions.

2.3 Growing awareness among operators

There is also growing awareness among VOD operators that users are sensitive to the issue of personal data protection. A pioneer of the algorithmic approach to video distribution, Netflix was also the first in the sector to be faced with the seriousness of this risk when it was forced to cancel its second recommendation contest following a complaint filed by a subscriber. She claimed Netflix infringed on her privacy when the company supplied contestants of the game’s first edition with a massive data set on its users, making it possible to identify them and reveal intimate information through cross-referencing with other data. Not only was the contest cancelled, but the suit was also settled out of court with the company understandably concerned about negative fallout[10] should the case become publicized.

[8] CNIL, “Les données, muses et frontières de la création : lire, écouter, regarder et jouer à l’heure de la personnalisation”, cahier IP n°3, Oct. 2015

[9] “CNIL, “Les données, muses et frontières de la création : lire, écouter, regarder et jouer à l’heure de la personnalisation”, cahier IP n°3, Oct. 2015

[10] SINGEL (R.), “NetFlix Cancels Recommendation Contest After Privacy Lawsuit”, Wired, Dec. 2010. https://www.wired.com/2010/03/netflix-cancels-contest/

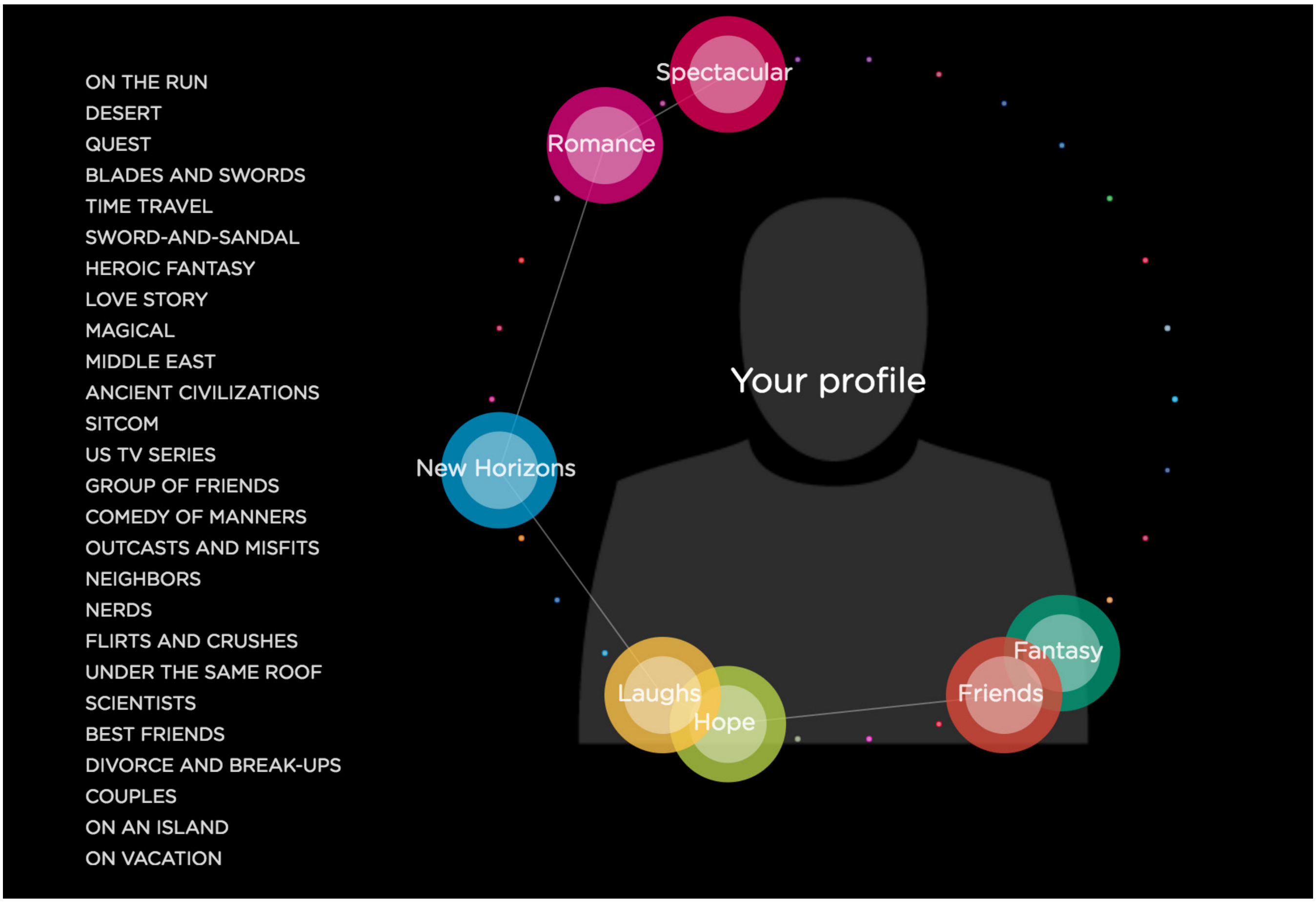
To the extent where personal data contains very precise information on a user's tastes and preferences, use of this data must go hand in hand not only with flawless data anonymization but also with particular care given to transparency.

This concern with respect of personal data is a point that is more and more important when video operators pick recommendation technologies offered by Spideo. Lucas Serralta, Digital Experience Director at Canal+ introduced in 2015 "Suggest", the new recommendation system powered by Spideo on CanaPlay, precisely through this transparency angle:

> *Our recommendation engine will function in real time and will tell the user why they have been recommended this content. We want them to understand the mechanism"*[11].

Franco-German TV channel Arte, frugal when it comes to user data, specifically prefers not to rely on gathering behavioral data to offer content recommendation. The channel explains its choice of Spideo technology through the prism of the proportionality principle. Alain Le Diberder, Director of Programs, declares:

> *We prefer a system where people indicate themselves what they prefer, instead of tracking them [...] Spideo's technology, which we are using, enables cross-referencing of formats and content which only seem to be unrelated on the surface but in fact are, such as suggesting to a user who just finished watched a documentary to watch a fiction next on a similar topic"*[12].



Dashboard of personal data visualisation

[11] Satellinet, 22 juin 2015, #250
[12] Satellinet, 22 juin 2015, #250

Approaches to content recommendation vary greatly from one video operator to the other. It is up to them to decide what kind of user experience they would like to promote and is related to their own editorial positioning. Semantic technologies developed by Spideo could take transparency as far as creating personal data visualisation dashboards for every user. When personal data is relevant, transparent and directly valuable for users, why not show it to them? We bet that in a near future, gaining the user's trust will require more than just opt-in / opt-out options.

The responsibility for recommendation suppliers such as Spideo is twofold. First, it is about offering tools that are modular enough so that each platform may build an experience that corresponds to it. Second, it is about giving the means to these distributors to use technologies as respectful as possible of personal data protection principles. Not only for ethical reasons but also, and probably especially, because these ethical choices are also winning choices when it comes to enabling new forms of digital TV to develop in the years to come through a long-term and trusting relationship between content providers and their users. ■

Conclusion

▶ Personal data is a sensitive topic, especially in the field of Recommendation & Analysis where data can tell a lot about individuals to the extent where it could be considered as breaching their privacy.

▶ Users are well aware of this and understandably concerned by potential breaches to the three guiding principles of data protection.

▶ Most recommendation technologies rely on using personal data only. From increasing financial sanctions to negative publicity and unsubscriptions, there is a risk in using personal data in the wrong way.

▶ Thankfully, not all recommendation technologies were created equal. Spideo's semantic technology focuses on knowing as much as possible on content in order to collect only relevant information from users.

▶ Keeping the best interest of users in mind is paramount. Transparency and making sure users understand why they are being recommended specific content is key to building lasting relationships and cementing loyalty.

▶ Data protection is anything but an obstacle for new media companies. By moving the focus away from users and towards content, it's a beneficial situation for users and content providers alike.

SPIDEO

SIMPLY RELEVANT

To learn more, please contact us :

Web : www.spideo.tv
Email : contact@spideo.tv
Twitter : @SpideoCorp
Tel : +33 9 81 92 82 99