Guillaume Cabanac, Max Chevalier, Claude Chrisment, and Christine Julien. Collective Annotation: Perspectives for Information Retrieval Improvement. In *RIAO'2007: Proceedings of the 8th conference on Information Retrieval and its Applications*. Electronic acts. May 2007.

Collective Annotation: Perspectives for Information Retrieval Improvement

Guillaume Cabanac,[1] Max Chevalier,[1,2] Claude Chrisment,[1] Christine Julien[1]
[1]IRIT/SIG – UMR 5505 CNRS
Université Toulouse 3 – 118 route de Narbonne
F-31062 Toulouse cedex 9

[2]LGC – ÉA 2043
IUT Rangueil – 129 avenue de Rangueil – BP 67701
F-31077 Toulouse cedex 4
{cabanac, chevalier, chrisment, julien}@irit.fr

Abstract

Nowadays we enter the Web 2.0 era where people's participation is a key principle. In this context, collective annotations enable to share and discuss readers' feedback with regard to digital documents. The results of this activity are going to be used in the Information Retrieval context, which already tends to harness similar collective contributions. In this paper, we propose a collective annotation model supporting feedback exchange through discussion threads. Considering this model, we associate annotations with a measure of the sparked consensus degree (social validation), this allows to provide a synthesized view of associated discussions. Finally, we investigate how Information Retrieval systems may benefit from the proposed model, thus taking advantage of human-contributed highly value-added information, namely collective annotations.

1 Introduction, Context & Motivation

Today we are waiting for the arrival of a new Web which is meant to be more open, as an evolution of the former "Web 1.0"; O'Reilly (2005) has named it "The Web 2.0." This new Web goes beyond its initial role of "a worldwide electronic library" as it shifts to being an alive, dynamic and interactive space where people may get involved and benefit from it. Web users go from a passive reader state to an active role of contributor. As a result, people's spontaneous contributions bring a real added value regarding the resources that they provide, describe or discuss. In this paper, we focus on a common activity allowing people to participate: collective annotation. Concerning Information Retrieval, one may notice that techniques are evolving with the emergence of this new user-centered Web. In this context, we propose a collective annotation model along with social validation that may contribute to Information Retrieval improvement.

Nowadays, people increasingly work with dematerialized digital documents: they gather, read and even write documents from many sources such as the Web. On paper, people usually pursue "active reading": they formulate marks and comments that support their critical thinking. As this activity is a key support for reading, it is transposed to digital libraries using "annotation systems" such as Annotea (Kahan et al., 2002). Gaining from modern computers capabilities, they enable networked readers to express their feedback by annotating documents, or even by annotating annotations themselves thus forming a discussion thread. Annotations along with their discussion threads have been shown to be useful. However, systems provide little information about their contents. In fact, understanding a discussion thread requires a high cognitive load because the reader has to read, understand and synthesize the arguments of each reply.

As a consequence, readers may not fully benefit from a widely annotated document. In order to overcome a such issue, this paper describes an approach for evaluating the "validity" of an annotation, i.e. the degree of consensus sparked by the given annotation. Since this validation is based on people's replies, we call this approach *social validation* of collective annotations. Thanks to this resulting information, readers are able to identify annotations that are globally confirmed (or refuted) by their discussion thread. Thus, they can focus on discussions that have ended up in a consensus; ongoing controversial discussions identification is also possible. As a result, the proposed approach may enable annotation systems to adapt annotations visualization according to users preferences, e.g. by emphasizing validated discussions. Regarding the Information Retrieval field, recent works tend to consider annotations as a source of evidence for better satisfying user requests, i.e. queries. We show that current approaches suffer from limits that may be overcome by taking into account the proposed collective annotation model along with social validation of annotations.

This paper is organized as follows: section 2 recalls how annotation activity has been practiced since the Middle Ages. The introduction of annotations into digital libraries tend to show that this activity is still needed nowadays. Regarding a survey of current annotation systems, we point out limits that may be overcome by the definition of a collective annotation model along with a social validation algorithm in section 3. Then, section 4 discusses how Information Retrieval may benefit from the identified proposed approach. Lastly, section 5 discusses strengths and weaknesses of our approach before concluding this paper and giving some insights into future research works in section 6.

2 From Paper to Digital Annotations

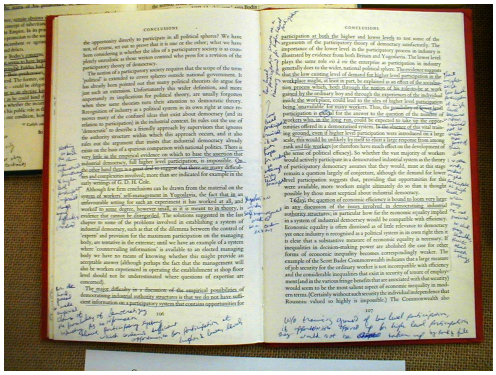
Annotating documents is an usual activity practiced since the early Middle Ages, e.g. Rabbi Rashi (1040–1105) is known for his comprehensive comments on the Talmud (Fraenkel and Klein, 1999). This activity traveled down the ages whatever the domain: Fermat's world-famous annotation written in about the year 1630 aroused more than 350 years of research in mathematics (Kleiner, 2000), notes on Blake or Keats' poems are still under study (Jackson, 2002) ... Still nowadays, Marshall (1998) shows that US undergraduates carry on annotating their textbooks. In fact, annotation is frequently used to support critical thinking and learning while reading, the whole activity being called "active reading" by Adler and van Doren (1972). In the following subsections, we present paper annotation virtues—for a personal and a collective use—as well as the transposition of this activity from paper to the digital documents. This leads us to describe the main features of software called "annotation systems."

2.1 Annotations Forms & Functions for Personal and Collective Use

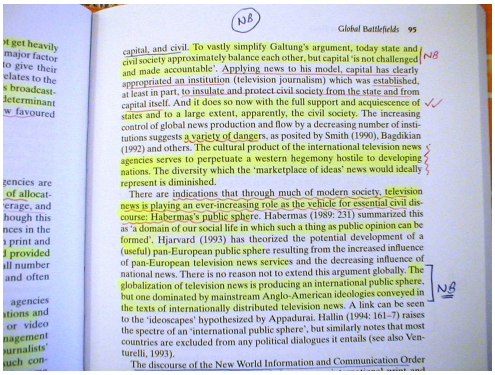
Numerous studies such as (Marshall, 1998; Ovsiannikov et al., 1999) have observed that readers mark documents expressing many styles by combining various tools (e.g. pencils, pens, highlighters, etc.) with colors. An annotation's location also reflects its function: a comment located in the header may sum up the document whereas an highlighted passage usually identifies an important piece of information to remember. This need to annotate and its results in terms of "marginalia" have been notably noticed by the Cambridge University Library.[1] Two representative kinds of annotations are depicted by figure 1: the left photo (a) shows comments in the

[1] cf. "Marginalia and other crimes" `http://www.lib.cam.ac.uk/marginalia`.

margins whereas the right one (b) represents highlight and emphasis marks like the '✓' symbol. We can also notice "idiosyncratic" marks: their semantics are only known by their author (e.g. ~~~~). Currently, annotations are mainly formulated for remembering, thinking through and clarifying text, for a personal use, according to Ovsiannikov et al. (1999). For a collective use, Wolfe and Neuwirth (2001) observed that they allow readers to provide feedback to writers or promote communication with collaborators as well as to address remarks directed to future readers while authoring. In past centuries, when books were rare and personal possessions,



(a)



(b)

Figure 1: Two representative annotated textbooks from the Cambridge University Library.

readers used to annotate and share their own copies (Jackson, 2002). Nowadays, annotating books borrowed from a library is considered defacement since they are public material. However, some people still seek for the "dirtiest," i.e. most annotated copy available because they consider previous readers' annotations valuable (Marshall, 1998). Now that documents are digitized, annotation value is reconsidered and introduced into the digital library through "annotation systems."

2.2 Annotation Systems : Transposing Paper Annotation Experience to Digital Libraries

Currently, documents tend to be drawn up using word processing software; they are mainly spread through networked computers. Murphy et al. (2003) noticed that reading documents on-screen is less comfortable, slower and less persuasive than reading paper. Moreover, readers feel frustrated of not being able to annotate digital documents, according to an experiment recounted by Sellen and Harper (2003, p. 95). Such inconveniences had already been pointed out by Ovsiannikov et al. (1999) who remarked that researchers preferred to print scientific papers so that they could annotate them while reading.

The need to annotate digital documents has soon been understood by researchers and companies that have been developing annotation systems since the early '1990s: consider for example Commentor (Röscheisen et al., 1994). Someone that is consulting a digital document can select a passage, create an annotation by using the appropriate function of the software and then type in a comment. Once the annotation is created,[2] it is generally displayed in-context, as close as

[2]Marshall (1998) qualifies such an annotation as "informal" since its structure and contents are subjective and not constrained. On the contrary, "formal annotations" such as metadata of the Semantic Web must be objective in essence and are mainly obtained from ontologies. In this paper, we only consider informal annotations and suggest readers to consult (Uren et al., 2006) for a recent state of the art about formal annotations.

possible to its anchor, i.e. the selected passage. See for example figure 2 which is a screenshot of the W3C Amaya (Kahan et al., 2002) annotation system where the characteristic yellow pen icon is inserted at the beginning of each annotated passage.

Some common software like Microsoft Word and Adobe Acrobat provide comment capabilities. Unfortunately, as comments are stored inside documents, they cannot be shared unless sharing documents themselves. On the other hand, most annotation systems store users' productions into dedicated annotation databases. In order to respect privacy, annotators can generally specify the scope of their annotations, e.g. private, restricted to specific users or public. When annotations are shared, they can be viewed by each user of the annotation system who may benefit from reading them. As the opinions expressed in annotations are subjective, people may recursively annotate them, forming a "discussion thread": a hierarchy of annotations ordered by their timestamps and rooted to its anchor, i.e. a document. Such a treelike organization of questions and answers dates back to Usenet forums introduced in 1979. For a further review of annotation systems technologies, we suggest referring to the following surveys (Ovsiannikov et al., 1999; Kahan et al., 2002; Wolfe, 2002).

2.3 Weaknesses of Current Annotation Systems

In the light of a survey of twenty annotation systems (Cabanac et al., 2005), this section identifies three common weaknesses. They concern systems provided by both commercial companies and research works. A first limit may concern annotation types that enable annotators to give further readers an overview of the annotation contents. In our view Amaya (Kahan et al., 2002) provides a comprehensive set of types {*Advice, Change, Comment, Example, Explanation, Question, SeeAlso*}. Annotators may associate at most one type with each annotation they formulate. As a result, when one wishes to express a question and an example within a single annotation, he is forced to choose a unique type (as they are exclusive) or he has to create two distinct annotations: one for each type. This constraint seems to be quite restrictive for numerous common situations. Indeed, one should be able to describe his annotations with any combination of types. In addition, aforementioned types allow the description of annotation contents only, e.g. thanks to *question* or *explanation* types. As a second limit, we may notice that such types do not carry information about the annotator's opinion. However, this information may be useful for further readers. Indeed, types such as "confirmation" and "refutation" would be informative for further readers. Moreover, such a hint would enable readers to get an overview of annotation contents without having to read its full contents, thus decreasing cognitive load.

The third limit concerns the scalability of annotation visualization in-context, i.e. within the annotated document. Indeed, a few annotations may be useful, therefore as a document is more and more annotated, it can be perceived as less and less usable (Wolfe and Neuwirth, 2001). For example, figure 2 shows what a widely annotated document may look like. The reader has no hint that would help him to understand their contents. Finally, numerous annotations may be compared to graffiti when they clutter the original text. Gaining advantage from annotations in figure 2 may be hard as they are numerous and people cannot spend time exploring each one. Indeed, one should remember that each annotation may have sparked a discussion thread. In this context, Wolfe and Neuwirth (2001) indicate that people may wish to identify controversy and consensus points. Then, in order to globally understand people's opinion with respect to the annotated passage, one has to read each item of the discussion, mentally extract their main

opinion and synthesize it. This requires a significant cognitive load that distracts the reader from his main task: reading. However, such a mental effort while reading should be avoided at all costs (O'Hara and Sellen, 1997). A first step towards reducing the profusion of inaccurate or unsuitable contents (e.g. advertisement, pornography) has been proposed by the JotBot system (Vasudevan and Palmer, 1999) which associates a lifetime with annotations. One may extend it by voting for an interesting annotation. This approach helps to eradicate graffiti. Therefore we suppose that interesting annotations get lost because they are located on resource having little audience. Moreover, users may not make the effort to vote for an annotation once they know its contents: this is a motivational issue.

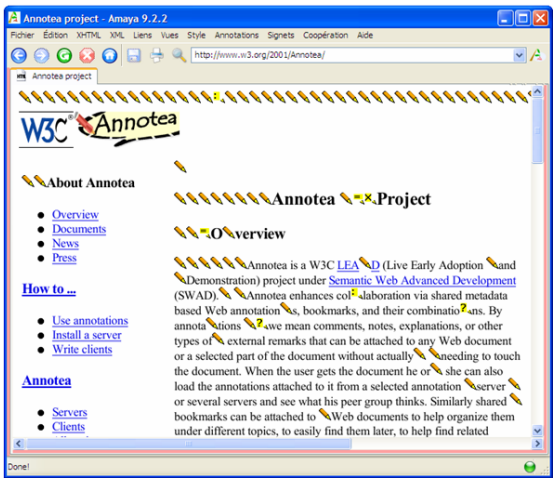


Figure 2: Illustration of the scalability limit concerning annotation in-context visualization.

To sum up, identified limits concern: *i*) the restriction to a maximum of one type per annotation; *ii*) the fact that types only characterize contents without reflecting author's opinions such as "confirmation" and "refutation"; *iii*) the scalability problem regarding visualization, i.e. a reader has no information about how much an annotation is validated by its discussion thread. In order to overcome these three limits, the following section defines the proposed approach aiming to measure the social validation of collective annotations.

3 Social Validation of Collective Annotations

In this paper, following Goguen's (1997) "Social Theory of Information," an annotation makes sense when a social group judges that it effectively makes sense. In our context, we consider that judgments are expressed within the discussion thread of the annotation; thus annotators that formulate replies constitute the aforementioned social group. To sum up, we qualify an annotation as *socially validated* when it is characterized by a discussion thread where people express a global consensus. This section first defines the concepts of *annotation* and *discussion thread* before describing algorithm that compute their "social validation."

3.1 Modeling Collective Annotations Along With Discussion Threads

An *annotation* (definition 1) is formulated by a user on a resource, e.g. a text, a picture. Concerning its location, anchoring points may be specified within the resource or on an already existing annotation. In the latter case, *replying* to it forms a *discussion thread* (definition 2). This section defines these concepts and illustrates them with a concrete example, see figure 3.

Definition 1. We define an *annotation* as a collective object that can be used for argumentative purposes. The "collective" adjective refers to the fact that an annotation may be consulted by any user of the annotation system. The "argumentative" facet of an annotation is because when annotating a resource, one may confirm or refute its contents. We model such an annotation as a pair $\langle OD, SI \rangle$ where:

1. The OD part represents Objective Data created by the annotation system. This consists of the following mandatory attributes for an annotation:
   - Its *identification* by a unique identifier such as a URI.
   - Its *author's identity* which enables to get information about him, e.g. name, email.
   - Its *timestamp* which enables to chronologically organize discussion threads.
   - Its *anchoring points* which unambiguously specifies its locations within the annotated resource. Various approaches have been published for semi-structured documents, e.g. Annotea (Kahan et al., 2002) relies on XPointer (DeRose et al., 2002).
2. The SI part represents Subjective Information formulated by the user when he creates an annotation. This consists of the following optional attributes for an annotation:
   - Its *contents*, a textual comment for instance.
   - Its *visibility* stored for privacy concerns, e.g. private, public, specified users.
   - Its author's *expertise* which may be useful for further readers as people tend to trust experts more than novices (Marshall, 1998).
   - The list of *references* provided by the annotator in order to justify its opinion. A book reference, a citation, a URL ... may be provided for that purpose.
   - Various *annotation types* that extend the works of Kahan et al. (2002) by providing opinion types. We have grouped them into two classes represented in table 1. Annotation types reflect the semantics of the contents (modification, example, question) as well as the semantics of the author's opinion (refute or confirm). By combining these types, authors may express suggestive point of views that are gradual, e.g. consider a $\mathcal{R}$-typed *versus* $\mathcal{RE}$-typed annotation. Concretely, these types may be inferred from annotation contents by Natural Language Processing (Pang et al., 2002) and then validated by their author.

Class name	Comment			Opinion (exclusive)	
Type name	modification	question	example	confirmation	refutation
Representation	$\mathcal{M}$	$\mathcal{Q}$	$\mathcal{E}$	$\mathcal{C}$	$\mathcal{R}$

Table 1: Annotation types for a collective annotation.

Definition 2. A *discussion thread* is a tree rooted on an annotation. This specific root annotation may be recursively annotated by *replies*. Note that replies are represented in a chronological order using their timestamps, i.e. their creation date and time.

Example 1. *Figure 3 represents a discussion thread where the contents of the refuting a annotation and its replies r_i are given in table 2. Note that $x \leftarrow y$ means that y confirms a part of x. Conversely $x \not\leftarrow y$ means that y refutes a part of x. Moreover, the hierarchy is constrained by annotation timestamps, e.g. $timestamp(r_1) \leqslant timestamp(r_2) \leqslant timestamp(r_3)$.*

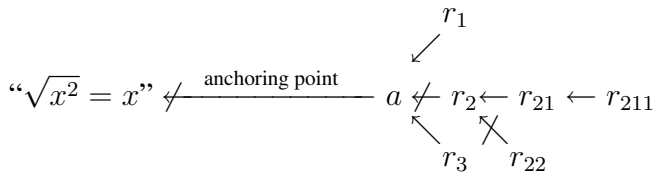


Figure 3: Discussion thread about a mathematical formula.

Annotation	Types	Annotator's comments
a	$\mathcal{RME}$	This formula is wrong, check this counterexample: $\sqrt{(-2)^2} \neq -2$ Consider the following modification: $\sqrt{x^2} = \|x\|$.
r_1	$\mathcal{CE}$	OK, for example $\sqrt{(-4)^2} = \|-4\| = 4$.
r_2	$\mathcal{R}$	This high school lesson only considers positive numbers, see part 2.
r_3	$\mathcal{C}$	More generally $\forall (x, n) \in \mathbb{R} \times \mathbb{R}^* \quad \sqrt[n]{x^n} = \|x\|$.
r_{21}	$\mathcal{CM}$	Then precise $\forall x \in \mathbb{R}_+ \quad \sqrt{x^2} = x$.
r_{22}	$\mathcal{RE}$	May be confusing when quickly and superficially read!
r_{211}	$\mathcal{CM}$	$\mathbb{R}$ is unknown in high school: you should use "positive numbers" instead.

Table 2: Arguments associated with the mathematical discussion of figure 3.

3.2 Computing Social Validation of Collective Annotations

This section describes a set of algorithms that compute the social validity $v(a) \in [-1; 1]$ of an annotation a. This continuous function reflects the intrinsic opinion of a as well as the global opinion expressed by its replies. Note that opinions are defined in table 1. Regarding its interpretation $v(a) \to 0$ means either that a has no reply or that replies expressing confirmation and refutation are balanced. Moreover $v(a) \to 1$ indicates that a is totally confirmed by its replies; conversely $v(a) \to -1$ means that a is totally refuted by its replies. As a result a consensus (either refuting of confirming) is identified when $|v(a)| \to 1$. In order to define how $v(a)$ evolves, we consider the combination of opinion values for the annotation and its replies: table 3 describes the four possible cases. For instance, Case 2 shows that replies that globally refute ($\mathcal{R}$) an annotation a (which is $\mathcal{C}$-typed) makes its social validity $v(a)$ decreasing: $v(a) \to 0$.

	Case 1	Case 2	Case 3	Case 4
Opinion of the parent annotation a	$\mathcal{C}$	$\mathcal{C}$	$\mathcal{R}$	$\mathcal{R}$
Global opinion of replies to a	$\mathcal{C}$	$\mathcal{R}$	$\mathcal{C}$	$\mathcal{R}$
Social validity $v(a)$	$v(a) \to 1$	$v(a) \to 0$	$v(a) \to -1$	$v(a) \to 0$

Table 3: Social validity of a parent annotation regarding the opinion types of its replies.

In order to compute $v(a)$, we have explored three distinct approaches. The first one considers the κ coefficient (Fleiss et al., 2003) provided by social sciences. Assuming that κ is not suitable for our concern, the second approach is based on an empirical recursive scoring algorithm (Cabanac et al., 2005). As a third approach, we intend to ground social validation into a more formal setting. Therefore, we extend the Bipolar Argumentation Framework proposed by Cayrol and Lagasquie-Schiex (2005a,b) in order to compute social validation of collective annotations.

3.2.1 First Approach: Statistical Measure of Inter-rater Agreement

The Cohen's kappa coefficient $\kappa \in [0, 1]$ measures the agreement among $n = 2$ "raters" (persons providing a rating) which split up N items in k mutually exclusive categories (Cohen, 1960). A generalization of this coefficient called Fleiss' κ may be applied for $n \geqslant 2$ raters (Fleiss, 1971). The value of the kappa coefficient $\kappa = \frac{P(A)-P(E)}{1-P(E)}$ varies according to $P(A)$, the agreement among the n raters as well as $P(E)$ representing chance agreement. Different ranges of values for κ have been characterized with respect to the degree of agreement they suggest. Concerning content analysis, a standard methodology in the social sciences, a value such that $\kappa > 0.8$ indicates a good degree of agreement (Krippendorff, 1980).

Using κ for measuring inter-annotator agreement would require that n annotators associate N annotations with a combination of types among the $k = 24$ available ones.[3] Thus, in order to compute the κ coefficient, the n annotators should have typed the N annotations by themselves. This is not realistic in the context of an annotation system that may manage hundreds of annotations. Moreover, this coefficient does not take into account the fact that a rating may be disputed by other raters, thus forming a discussion tree. All things considered, a κ-based approach does not seem to be adapted for solving our concern. As a discussion thread is a tree of annotations, the following section proposes a recursive scoring algorithm as a second approach.

3.2.2 Second Approach: Empirical Recursive Scoring Algorithm

In a second attempt, we have defined a recursive algorithm that computes the social validation of an annotation (Cabanac et al., 2005). This algorithm may be viewed as a twofold process working on the annotations of a discussion thread, i.e. the root annotation and its replies. The first step consists of evaluating the intrinsic *agreement* of each annotation using two parameters. *i*) The *confirmation* value $c(a)$ of an annotation a is based on its opinion types. Indeed, the confirmation $\mathcal{C}$-typed implies a positive confirmation value whereas the refutation $\mathcal{R}$-typed implies a negative confirmation value. Graduality in confirmation evaluation is obtained by considering combination of types, see figure 4.

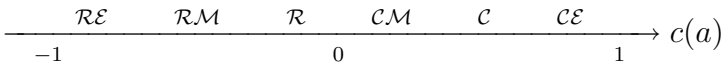


Figure 4: *Confirmation* value of an annotation a regarding its types.

ii) We assume that the *agreement* of an annotation has to be increased considering the involvement of its author. Indeed, when someone provides a comment as well as some references (i.e. optional information) then he makes an extra cognitive effort in order to justify his opinion. Such information brings an additional value: just compare a refutation only (e.g. "This is wrong!") with commented references that bring proof of the refutation (e.g. "Regarding X and as it can be seen on Y, I think that it is false because ..."). Concretely, if A denotes the set of annotations, the $i_c : A \rightarrow [0, 1]$ function (1) reflects the presence of a comment whereas the $i_r : A \rightarrow [0, 1]$ function (2) varies according to the number of provided references. Note that the dotted notation refers to the annotation model of definition 1, e.g. $a.contents$ means "the contents of the a annotation."

[3] $k = |\{\mathcal{R}; \mathcal{C}; \varnothing\} \times \{\mathcal{MQE}; \mathcal{MQ}; \mathcal{ME}; \mathcal{QE}; \mathcal{M}; \mathcal{Q}; \mathcal{E}; \varnothing\}| = 24$ is the number of possible combinations of the two annotation classes (comment and opinion) that are described in table 1.

$$i_c(a) = \begin{cases} 0 & \text{if } ||a.contents|| = 0 \\ 1 & \text{else} \end{cases} \quad (1) \qquad i_r(a) = \frac{|a.references|}{1 + \max_{x \in A} |x.references|} \quad (2)$$

The *agreement* of an annotation is evaluated by the $a : A \rightarrow [0, 1]$ function (3) that takes into account the types as well as the contents (comment, references) and the confirmation value $c(a)$ of an annotation. The $\alpha, \beta \in [0, 1]$ parameters allow to adjust the weight of these two parameters.

$$a(a) = \frac{c(a)\ (1 + \alpha \cdot i_c(a))\ (1 + \beta \cdot i_r(a))}{(1 + \alpha)\ (1 + \beta)} \quad (3)$$

The second step consists of aggregating the intrinsic agreement of the annotation with the global agreement expressed by its replies. This aggregation called *social validation* is computed by the $v : A \rightarrow [-1, 1]$ function (4). In other way, $v(a) = 0$ if a is the root ($a.ancestor = \lambda$) of an empty discussion thread ($|a.replies| = 0$) which means that a is neither confirmed neither refuted. Otherwise, $v(a)$ is computed according to its *agreement* value and a *synthesis* of its replies. To do that, the s function returns a negative value when the replies globally refute the parent item and a positive value otherwise. Finally, the $\gamma \in [0, 1]$ parameter allows to adjustment of the impact of the discussion thread on the measure of the social validity of a.

$$v(a) = \begin{cases} 0 & \text{if } a.ancestor = \lambda \wedge |a.replies| = 0 \\ \frac{1}{2} \cdot a(a) \cdot (1 + \gamma \cdot s(a)) & \text{else} \end{cases} \quad (4)$$

The *synthesis* $s : A \rightarrow [-1, 1]$ function (5) is important because it takes into account the replies of an annotation as a whole. We base this function on a weighted mean that gives a prominent role to replies qualified with a greater expertise (the range of the expertise attribute being strictly positive). We also increase the value of the *synthesis* considering the number of replies: the more an annotation has replies, the more its weight is taken into account.

$$s(a) = \begin{cases} 1/\gamma & \text{if } |a.replies| = 0 \\ \frac{\sum_{r \in a.replies} v(r) \cdot r.expertise}{\sum_{r \in a.replies} r.expertise} \left[1 + \ln\left(\frac{|a.replies|}{m(1+l(a))}\right) - \ln(2)\right] & \text{else} \end{cases} \quad (5)$$

In equation (5), $a.replies$ denotes the set of replies associated to the a annotation. Moreover, the $l : A \rightarrow \mathbb{N}_+$ function returns the level of a given annotation in the discussion thread, the root annotation being at level 0. Finally, the $m : \mathbb{N}_+ \rightarrow \mathbb{N}_+$ function returns for $m(l)$ the maximum number of replies at the l^{th} level of the tree corresponding to the discussion thread.

To sum up, if an annotation a belongs to the discussion thread and does not have any reply, its *social validity* correspond to the *agreement* value: $|a.replies| = 0 \Rightarrow v(a) = a(a)$. On the contrary, if a has got replies, the *synthesis* values relies on the weighted mean of their social validity (A) multiplied by an expression (B) increasing with the number of replies. B takes into account the maximum number of replies existing for the same level than the level of a. For instance, if a (level l) has n replies and if the maximum number of replies for the annotations at the l^{th} level is N, then $B = 1 + \ln(1 + \frac{n}{N}) - \ln(2)$. Note that the upper bound of B is reached when $n = N \Rightarrow B = 1$. Therefore $B \in [\ln(\frac{e \cdot (N+1)}{2N}), 1]$, we use a natural logarithm for reducing differences between small and large values of N. As $B \leqslant 1$, it cannot increase the value of A.

The $v(a)$ value of the *social validation* associated with the a annotation allows to evaluate it according to the *agreement* and *synthesis* functions. Indeed, $|v(a)| \to 1$ points out an annotation a that contains replies that globally agree with it. According to the sign $v(a)$, one can conclude that people validate a confirm $\mathcal{C}$-typed (resp. a refute $\mathcal{R}$-typed) annotation when $v(a) \to 1$ (resp. $v(a) \to -1$).

The social validation algorithm presented in this section is mostly based on heuristics and empirically instantiated parameters. In order to go beyond this first empirical approach, the following section describes another proposition for computing social validation. This approach is grounded on formal research by Cayrol and Lagasquie-Schiex (2005a,b) from the Artificial Intelligence field. Its main advantage lies in the fact that it takes into account argumentation theoretical principles, moreover it has been used in numerous contexts, e.g. medical decision making.

3.2.3 Third Approach: Extending a Formal Bipolar Argumentation Framework

Cayrol and Lagasquie-Schiex's Bipolar Argumentation Framework is based on Dung's (1995) argumentation framework. The latter is defined by the pair $\langle A, R \rangle$ where A is a set of arguments and R is a binary relation on A^2 called a defeat relation. Thus, an argumentation framework can be represented as a directed graph whose vertices are arguments and edges represent their targets. Identifying attack and defense branches allows us to decide on an the acceptability of an argument—which is a binary value: acceptable or not acceptable—regarding conflict-free and collective defense sets. In addition, (Cayrol and Lagasquie-Schiex, 2005a) notice that most work on argumentation only consider a single type of interaction between arguments: the *attack*. However, previous works (Karacapilidis and Papadias, 2001) have suggested that, for numerous concrete contexts, another type of interaction must be considered in order to realistically represent knowledge. This other type of interaction is *defense*. Thus, regarding attack and defense arguments, Cayrol and Lagasquie-Schiex extend Dung's (1995) argumentation framework in order to define a Bipolar Argumentation Framework.

Definition 3. A *Bipolar Argumentation Framework* (BAF) is a triplet $\langle A, R_{app}, R_{att} \rangle$ where:

- $A = \{a_1, \ldots, a_n\}$ is a set of arguments,
- R_{app} is a binary *support* relation on A^2. The pair $(a_i, a_j) \in R_{app}$ is denoted by $a_i \to a_j$,
- R_{att} is a binary *defeat* relation on A^2. The pair $(a_i, a_j) \in R_{att}$ is denoted by $a_i \not\to a_j$.

Besides definition 3, Cayrol and Lagasquie-Schiex (2005a) define a gradual valuation v on their framework, following 3 principles. **P1**: the valuation of an argument depends on the values of its direct defeaters and of its direct supporters. **P2**: if the quality of the support (resp. defeat) increases then the value of the argument increases (resp. decreases). **P3**: if the quality of the supports (resp. defeats) increases then the quality of the support (resp. defeat) increases. Regarding these principles, the authors set $a \in A$ with[4] $R^-_{app}(a) = \{b_1, \ldots, b_p\}$ and $R^-_{att}(a) = \{c_1, \ldots, c_q\}$. Finally, they define a gradual valuation as the application (6) $v : A \to V$ such that:[5]

$$v(a) = g(h_{app}(v(b_1), \ldots, v(b_p)), h_{att}(v(c_1), \ldots, v(c_q))) \tag{6}$$

[4] $R^-_{app}(a)$ (resp. $R^-_{att}(a)$) are direct supports (resp. defeats) of the a argument.

[5] V denotes a completely ordered set with V_{min} a minimum element and V_{max} a maximum element. Moreover, V^* denotes the set of finite sequences of element V including the empty sequence.

with the function h_{app} (resp. h_{att}) : $V^* \rightarrow H_{app}$ (resp. $V^* \rightarrow H_{att}$) that evaluates support (resp. attack) upon an argument. The function $g : H_{app} \times H_{att} \rightarrow V$ with $g(x, y)$ is increasing on x and decreasing on y. The function h ($h = h_{app}$ ou h_{att}) must satisfy the 3 following conditions: **C1** if $x_i \geqslant x'_i$ then $h(x_1, \ldots, x_i, \ldots, x_n) \geqslant h(x_1, \ldots, x'_i, \ldots, x_n)$; **C2** $h(x_1, \ldots, x_i, \ldots, x_n, x_{n+1}) \geqslant h(x_1, \ldots, x_i, \ldots, x_n)$; **C3** $h() = \alpha \leqslant h(x_1, \ldots, x_i, \ldots, x_n) \leqslant \beta$ for all $x_1, \ldots, x_i, \ldots, x_n$.

The authors propose two instances for this generic evaluation. In the first, argument values are aggregated by retaining the maximum of direct attacks and supports, i.e. $h_{att} = h_{app} = \max$. This fist approach is not acceptable for our context because it does not take into account the whole set of expressed arguments. In the second approach an instance is parameterized by $V = [-1, 1]$, $H_{app} = H_{att} = [0, \infty]$, $h_{app} = h_{att} = \sum_{i=1}^{n} \frac{x_i+1}{2}$ et $g(x, y) = \frac{1}{1+y} - \frac{1}{1+x}$.

We have chosen to retain the latter evaluation instance for our context. In order to socially validate a collective annotation, we model its discussion thread by a BAF. Its A set contains nodes of the discussion thread; pairs of the R_{app} (resp. R_{att}) set are defined by annotations and reactions of the confirm $\mathcal{C}$ (resp. refute $\mathcal{R}$) type along with their parent targets.

Example 2. *The discussion of figure 3 is modeled by $\langle A, R_{app}, R_{att} \rangle$, a Bipolar Argumentation Framework where $A = \{a, r_1, r_2, r_3, r_{21}, r_{22}, r_{211}\}$ figures the annotation a and its replies r_i. Relations between annotations are expressed by $r_{app} = \{(r_1, a), (r_3, a), (r_{21}, r_2), (r_{211}, r_{21})\}$ for the $\mathcal{C}$ type whereas $r_{att} = \{(r_2, a), (r_{22}, r_2)\}$ reflects the $\mathcal{R}$ type.*

The gradual evaluation $v(A) = 0.152$ of this example does not take into account some available argument data. Indeed some subjective information (SI, cf. definition 1) associated with nodes of the discussion thread are not considered, e.g. "comment" class types, expertise, comment and references. This is the reason we extend the BAF by redefining the v application (6). Thus, we provide the $v' : A \rightarrow V$ application (7) such that:

$$v'(a) = g(h_{app}(i(b_1) \times v'(b_1), \ldots, i(b_p) \times v'(b_p)), \\ h_{att}(i(c_1) \times v'(c_1), \ldots, i(c_q) \times v'(c_q))) \tag{7}$$

We introduce the $i : A \rightarrow I$ function (8) as a parameter of $h_{app} : V^* \rightarrow H_{app}$ and $h_{att} : V^* \rightarrow H_{att}$ in order to measure the intrinsic value of an argument by taking into account n criteria. Our choice for the second evaluation instance requires that $V = [-1, 1]$. Therefore, we define $I = [0, 1]$ in order to respect $\forall a \in A \quad i(a) \cdot v'(a) \in V$. Moreover, the n coefficients $\pi_i \in [0, 1]$ defined such that $\sum_{i=1}^{n} \pi_i = 1$ allows adjustment of the relative importance of the n criteria evaluated by $f_i : A \rightarrow F_i \subseteq \mathbb{R}_+$ functions. Note that the $\sup$ function returns the least upper bound of the f_i functions domain of definition. The $\delta \in [0, 1]$ coefficient allows adjustment of the n criteria global impact on the v' evaluation, note that $\delta = 0 \Longrightarrow v'(A) = v(A)$.

$$i(x) = \delta \cdot \sum_{i=1}^{n} \frac{\pi_i \cdot f_i(x)}{\sup(F_i)} \tag{8}$$

In our applicative context, we have identified $n = 4$ criteria that may be taken into account to evaluate an argument: expertise and agreement of its supporters and defeaters as well as it's author's implication in terms of comments and references. Thus, the f_1 function increases according to the *expertise* associated with the evaluated argument; we propose a five-point expertise

scale : novice $\prec$ beginner $\prec$ intermediary $\prec$ confirmed $\prec$ expert. The f_2 function associates a real value representing the annotator's agreement to each combination of types from the "comment" class. For example, somebody who gives an example ($\mathcal{E}$) agrees more than somebody who proposes a modification ($\mathcal{M}$).

$$\xrightarrow[F_2]{\mathcal{MQE} \quad \mathcal{MQ} \quad \mathcal{ME} \quad \mathcal{QE} \quad \mathcal{M} \quad \mathcal{Q} \quad \mathcal{E}} f_2(a)$$

The f_3 function evaluates the annotator's implication regarding the existence of a *comment* on a given annotation. Finally, the f_4 function increases with the number of *references* cited by an annotation: $f_4(x)$ is the ratio between the number of references contained by x and the maximum number of references per annotation. Regarding implementation of the proposed approach, we have fixed $\delta = 1$ so that the impact of the n criteria on the evaluation of arguments to be maximal. While foreseeing to experiment other weightings, we intuitively define $\pi_2 = \pi_3 = \frac{1}{3}$ and $\pi_1 = \pi_4 = \frac{1}{6}$ so that it mostly takes into account comments as well as agreement of attacks and supports.

An annotation is the subject of social consensus, i.e. it is *socially validated* when its reactions globally express a single opinion, either refutation or confirmation. This is obtained when $(v'(A) \to 1) \vee (v'(A) \to -1) \iff |v'(A)| \to 1$. These results really represent a semantical-based, in-context synthesis of social groups' opinions. The next section intends to investigate how Information Retrieval and Visualization related techniques may benefit from such valuable information.

4 Collective Annotations: Perspectives for Information Retrieval Improvement

Since annotations on documents provide human-contributed information, researchers in the Information Retrieval field have proposed to exploit them in order to improve annotation-based document retrieval. Section 4.1 presents the main published strategies for that purpose. Then, we identify limits of these current strategies in section 4.2. In our view, the outlined limits may be dealt with due to the annotation model and *social validation* approach proposed in this paper.

4.1 Improving Information Retrieval Through Collective Annotations: Current Works

Since the early 1960s, Fraenkel has been building the Responsa Retrieval Project which aims at storing full-text corpora and providing advanced search capabilities (Choueka, 2001). The corpora exploited by this project comprises 400 annotated ancient books—including the Bible and the Talmud—representing altogether 172 million words.[6] Fraenkel and Klein (1999) state that annotations can improve document retrieval since an annotator's comments generally explain or discuss a passage in his own words. It is particularly the case for the Talmud which is "extremely condensed and often impossible to understand without the help of the many layers of commentaries which have been added over the centuries." Therefore, considering such human-contributed comments may lead to improved document recall.

In order to enable people to find documents using their related annotations, Fraenkel and Klein (1999) explored the following two alternatives. *i*) Embedding annotations into the regular text is the easiest way to proceed, requiring little overhead. However, this strategy does not allow a

[6] These statistics come from the Responsa Retrieval Project web site: `http://www.biu.ac.il/JH/Responsa`.

query to target either documents or annotations alone, neither to restrain a search on some annotators' comments, for example. Moreover, disrupting the main text by inserting annotations into it may cause wrong results, e.g. exact phrase search such as "w_1 w_2" would wrongly return documents containing the w_1 word followed by any annotation starting by the w_2 word. *ii*) Treating annotation as special text requires additional storage and processing efforts. Though, keeping a boundary between original documents and their annotations enables to overcome aforementioned limits. That is the reason why the second approach is preferred.

More recently, in the Digital Library research field, research works have tackled annotated document retrieval again. These works consider the second alternative as annotations and documents are generally separately stored in a Digital Library. Indeed, Agosti and Ferro (2005, 2006) view annotations as sources of evidence: they can be used for retrieving more documents as well as to re-ranking them. Globally, their proposed algorithm takes the user query q as input and builds the corresponding document result list as follows. First, the system retrieves the ranked list D of documents matching q. Second, it retrieves the ranked list A of annotations whose contents match q. Note that elements of D and A are ranked by computing a similarity between the query and element contents only. Third, it determines the set D' of documents related to A. Fourth, the result provided to the user is a combination of D and D'. The latter point uses traditional data fusion techniques (Lee, 1997). Concretely, Agosti and Ferro (2005) give an example that illustrates why taking into account annotations for direct search may be worthwhile. They consider a query $q =$"`good survey grid computing`." The matching documents $D = \{d_3, d_4\}$ are first retrieved because they contain at least one term of the query. Moreover, matching annotations $A = \{a_6, a_7, a_{12}\}$ are also found for similar reasons: a_7 contains "`good survey`," for example. Considering A as a source of evidence that helps in finding more documents, the corresponding documents $D' = \{d_2, d_3\}$ are then considered as relevant. Indeed, d_3 is present in both D and D' which makes its score greater than the one of d_2 which is only in D'. Moreover, d_2 would have stayed undiscovered without querying the annotations.

Thiel et al. (2004) have also considered annotation for Information Retrieval in the context of the COLLATE[7] project. In this setting, scientists had to interpret and index digitized material (such as press articles, photos, films fragments) thanks to annotations along with discussion threads. In order to characterize the aim of an annotation, types related to interpretation (e.g. comparison, cause, background information) as well as argumentation (support argument, counterargument) are provided. For previously reasons, annotations are considered for IR. This facet of COLLATE is described in (Frommholz et al., 2003): a document result list retrieved with respect to a given query is re-ranked using an empirical recursive algorithm.

Finally, Frommholz et al. (2004) formalize a logic-based retrieval approach using a probabilistic datalog, i.e. a framework based on probabilistic logics. This collective annotation model is refined and experimented in (Frommholz, 2005) on the TREC enterprise track[8] (Voorhees and Buckland, 2005) that consists of two tasks: "Expert search" and "Email search." For the latter task, Frommholz model email threads as discussion threads, see (Frommholz and Fuhr, 2006). In the following section, we compare current works with the proposed approach.

[7] The EU-funded "Collaboratory for Annotation, Indexing and Retrieval of Digitized Historical Archive Material" project `http://www.collate.de`.

[8] cf. the TREC Enterprise task `http://www.ins.cwi.nl/projects/trec-ent`.

4.2 Discussion: Perspectives for Collective Annotations With Respect To Current Works

Section 4.1 has provided an overview of recent literature on annotated document retrieval. Regarding each contribution in detail, section 4.2.1 identifies some limits. Then, we argue in section 4.2.2 that the proposed collective annotation model along with social validation may represent an initial contribution to overcome these limits. Finally, we outline some exploratory paths regarding how Objective Data brought by annotations may help to improve information retrieval on annotated corpora.

4.2.1 Limits of Current Information Retrieval Works Based on Collective Annotations

A first limit may concern systems that do not distinguish positive from negative annotations, e.g. FAST (Agosti and Ferro, 2006) or Amaya (Kahan et al., 2002). Indeed, an annotation expressing "a good survey" on a d_1 document is considered just the same as another which expresses "not a good survey" on d_2. As a result, d_1 and d_2 are both retrieved when querying for "good survey." However, d_2 should not be retrieved as it represents the opposite of user needs formulated by the query. This example illustrates limits of retrieval that is exclusively based on query-annotation contents similarity. Some systems get rid of this limit by providing annotators with the "confirmation" and "refutation" argumentative types, cf. COLLATE (Thiel et al., 2004) or POLAR (Frommholz and Fuhr, 2006) for instance. This introduces valuable semantics, yet a limit remains because such types cannot express fine-grained opinions, i.e. graduality in argumentation that is needed in real situations: consider "quite sure" *versus* "absolutely sure."

As a second limit, we wonder why the richness of annotation contents is actually poorly exploited by current approaches. For example, an annotator may include examples, counterexamples, proofs, references, citations, etc. in his annotation contents. Indeed, these elements may suggest to further readers that this annotation should be more reliable than another but that only expresses an individual opinion. In addition, people may have different areas of expertise: one may be a novice in a domain whereas being a recognized expert in another—consider how expertise is taken into account by conference program committees. To the best of our knowledge, annotation systems do not consider these valuable elements that are called Subjective Information in the proposed annotation model, see definition 1.

Third, in previously research works, annotations enable search engines to find more documents. A document may be considered as relevant either because its contents or the contents of its annotations match the query. Currently, a user viewing a result list cannot understand the ranking of a specific document: has it been retrieved because of its content, its annotations or both of them? In addition, when viewing a given document later, systems do not provide information about people's reactions expressed within annotations (e.g. global refutation, weak confirmation, neutral).

4.2.2 Contributions of the Proposed Collective Annotation Model and Social Validation

In our opinion, systems may gain in accuracy if the above weaknesses are overcome. As a result, we discuss in this subsection how our approach may contribute to reducing these weaknesses.

Concerning the first limit (the need for gradual argumentation) the proposed *annotation model* (definition 1) provides two classes of annotations (cf. table 1). By combining types that belong to these classes, an annotator may express gradual opinions.

The second point concerns the limited use of annotation contents. To overcome this limit, our model $\langle OD, SI \rangle$ is composed of Objective Data as well as Subjective Information. This latter component clearly gathers distinct elements (contents, reference, expertise, etc.) that are used to adjust the *social validation* of each annotation.

Last but not least, the third issue concerns the understanding of how documents are retrieved. Annotation systems may make good use of social validation by adapting visualization to the degree of consensus achieved by each annotation. This proposition is introduced in our research prototype called TafAnnote (Cabanac et al., 2005, 2006). Concretely, it adapts the visual restitution of annotations by employing visual metaphors, i.e. a specific icon for each type. Moreover, it emphasizes *socially validated* annotations so that people may focus on discussions characterized by a social consensus, for instance.

Before further describing the TafAnnote prototype in section 5, we may propose original ways to improve collective annotation-based Information Retrieval. For example, referring to the proposed annotation model (definition 1) and considering existing works on annotation integration for Information Retrieval, one may notice that OD, i.e. Objective Data is not taken into account. In particular, we notice that annotation locations are not considered with respect to the user's query in an annotation-based retrieval context. However, we think that it is worthwhile considering the query-annotation similarity. This may be used for adjusting the impact of an annotation on its document weight. *Social validation* may also be considered as a factor that increases an annotation's impact. For example, the contents of a positive (i.e. confirming) annotation along with its discussion thread may be useful for enriching the annotated document passage as they are additional contents that have been qualified as trustworthy by a social group. On the contrary, a negative (i.e. refuting) annotation may be used for decreasing the importance of the passage with respect to the given query. For example, consider a passage which matches a user query. If it is refuted ($\mathcal{R}$-typed) by annotations, this means that it is perceived as somewhat dubious by a social group. That is why the retrieval system should reconsider its relevance.

In order to get an initial framework suitable for validating the proposed annotation model and social validation approach, we have developed the prototype presented in next section.

5 Implementation & Evaluation of the Proposed Approach

As a first step towards experimentation, we have developed an annotation prototype called TafAnnote (Cabanac et al., 2005) whose global architecture is illustrated by figure 5.

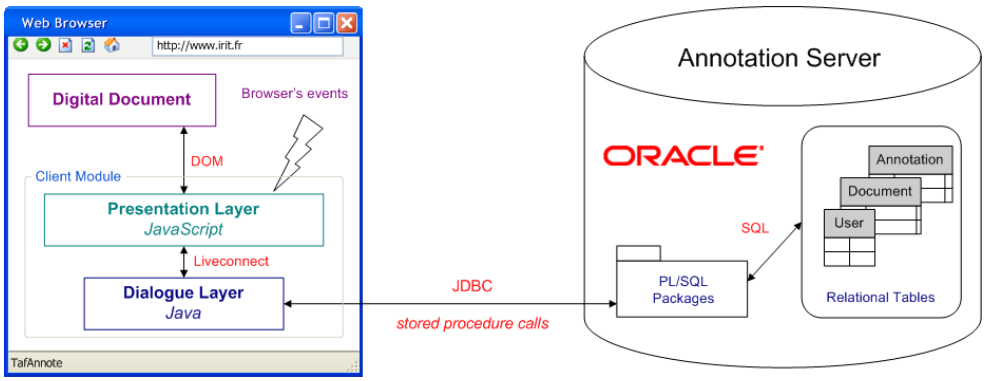


Figure 5: Global architecture of the TafAnnote (Cabanac et al., 2005) prototype.

In fact, TafAnnote relies on a twofold architecture based on several technologies. In one case, the client module developed in Javascript and Java is integrated in the Mozilla Firefox Web browser as an "extension." In the other, the annotation server is developed on top of an Oracle relational database. The interface between the two modules is achieved by stored PL/SQL procedure calls through the network. For example, when a user browses a Web page, this triggers an event which is caught by the client module. This latter component then queries the annotation server—which may be anywhere on the network—for any annotation about the given URL. When annotations are retrieved, the client module finally inserts annotations by manipulating the Document Object Model through the standard DOM API[9] provided by any modern Web browser. In the end, the user gets a similar visualization to figure 6.

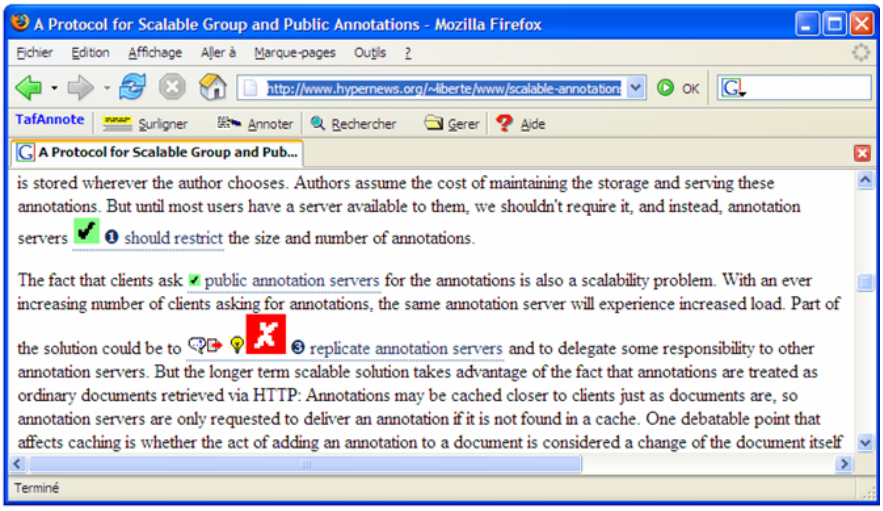


Figure 6: Visualizing annotations along with *social validation* in TafAnnote.

Annotation types in the proposed model (definition 1) are represented by visual metaphors, i.e. specific icons for each type. They are emphasized according to annotations' *social validation* values. In figure 6, the $\mathcal{R}$-typed annotation is emphasized because it is more socially validated than the two other $\mathcal{C}$-types others. This means that people agree with the $\mathcal{R}$-typed annotation and so it may be worth consulting it in order to obtain a different opinion to that of the document's author. The user obtains a visualization of the consulted annotation's discussion thread, see figure 7 (b). Compared to the visualization proposed by the Amaya (Kahan et al., 2002) Web browser provided by the W3C and INRIA, TafAnnote provides an overview of each reply by type. Combined with the *social validity* value of the root annotation, this may help people to identify consensus and controversy points within discussion threads.

The toolbar visible on figure 6 provides access to the system's functions: highlighting or annotating a passage,[10] searching for annotations, and viewing one's annotation hierarchy. The latter function enables a user to manage his annotations. At creation, they are stored in the user's personal hierarchy that he can reorganize by drag and drop. The system also provides an alternative visualization representing groups of annotations according to their types.

[9] The Document Object Model is a W3C recommendation, cf. `http://www.w3.org/DOM`.

[10] As noticed by Denoue and Vignollet (2000), users may wish to only highlight document passages without providing additional information, e.g. contents, references ... As a result, TafAnnote provides a such feature that creates "empty" annotations without preventing annotators to complete them later.

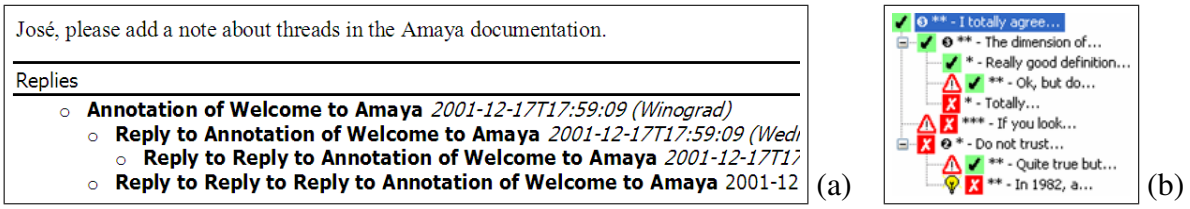


Figure 7: Visualizing a discussion thread in two annotation systems: (a) Amaya (Kahan et al., 2002) *versus* (b) the TafAnnote (Cabanac et al., 2005) prototype.

TafAnnote is an effective annotation system prototype actually designed for experimentation purpose. We plan to evaluate our approach, i.e. the annotation model and social validation with users. We want to evaluate how close *social validation* is to human perception of consensus and controversy. It should be worthwhile analyzing result differences, if any, between the empirical (cf. section 3.2.2) and formal (cf. section 3.2.3) proposed algorithms. Regarding experimentations related to annotation systems Wolfe and Neuwirth (2001) complain about protocols that only consider experts who use computers daily, e.g. researchers, engineers, undergraduate students. That is why we should investigate and confront results gained from experts as well as from "standard" users. Concerning the latter users, we have already noticed that people often misunderstand the representation of figure 7 (b). This especially happens when they are not used to reading discussion threads. In fact, people think that a reply's types are related to the root of the discussion thread, and yet it is related to its direct parent reply. In order to overcome this issue, we propose the alternative FreeMind-like[11] visualization of figure 8 that represents a discussion thread in a more "scattered" way.

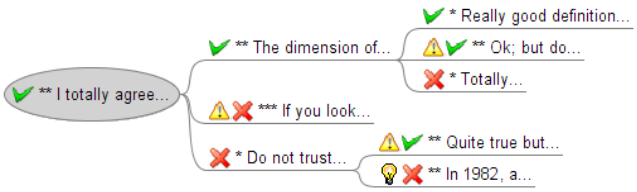


Figure 8: Proposed "scattered" view of the discussion thread represented in figure 7 (b).

Concerning the experimentation itself, we regret the lack of annotated test collections. This has already been mentioned by other researchers such as Agosti and Ferro (2006) and Frommholz and Fuhr (2006). Therefore we have to build a suitable collection in order to evaluate our approach. This is a long and difficult task as we have to obtain annotated documents along with discussion threads. Then, we may have to manually extract the attributes required by our model, e.g. expertise, references, types. As a first approach, we intend to derive an existing collection in order to experiment our approach. This may be achieved by using TREC email corpora (Voorhees and Buckland, 2005) provided for the "Enterprise" track, for instance. Indeed, Frommholz and Fuhr (2006) consider that email discussion threads are close to annotation ones.

6 Conclusion & Future Works

Annotating paper is a long practiced activity that helps people to think while they read. With the advent of the Web, researchers have made it possible to annotate digital documents. Such

[11] FreeMind is a mind-mapping software, cf. `http://freemind.sourceforge.net`.

systems such as Amaya (Kahan et al., 2002) are called "annotation systems." Thanks to modern computers and network capabilities, it has also been possible to share people's annotations in dedicated databases. This enables each user to consult previous readers' remarks that are expressed through annotations in the document context. Moreover, annotations may be annotated in turn as they are subjective contributions; this leads to create a discussion thread. In-context discussion threads may be very informative. However, their understanding requires reading of reply, extraction of its contribution and position regarding the replied resource and finally synthesizing of all the replies. In our view, this is tedious and highly cognitive. In order to reduce this cognitive load, this paper has described an approach based on a collective annotation model enabling *social validation* of annotations. This enables users to know the degree of consensus (either positive or negative) of a root annotation through the analysis of its associated discussion thread.

In current Information Retrieval research works, human-contributed information about specific document passages—namely informal annotations—are perceived as value-added material. Indeed, annotations are used to improve document retrieval. We have identified some limits on current works that may be overcome by our annotation model along with *social validation* of annotations. Moreover, we suggest that annotations' Objective Data—particularly anchoring points of annotations—should be exploited in order to make annotated document retrieval gain in accuracy. In order to experiment the proposed approach, we have developed the TafAnnote annotation system prototype. In the same way as many other researchers, we regret the lack of experimental test collections (Agosti and Ferro, 2006; Frommholz and Fuhr, 2006). We plan to derive an existing collection in order to experiment our approach. We thus intend to compare human perception of discussion thread consensus with the proposed empirical as well as formal algorithms that compute *social validation* of annotation. Considering discussion threads, we foresee the need to identify discussion drifts, i.e. when the topic of a discussion does not consider the root annotation anymore. This will enable the proposed *social validation* to gain in accuracy. Another future work will quantify how much Objective Data of the proposed annotation model—especially anchoring points—will improve annotation-based document retrieval.

Lastly, concerning the future of annotation we foresee that it may become the cornerstone of the ongoing "Web 2.0" (O'Reilly, 2005) based on people's participation. Currently, projects involving hundreds of thousands participants such as the Wikipedia encyclopedia[12] are interesting applicative contexts for the proposed annotation model and *social validation*. Indeed, wide participation implies a wide range of opinions that have to be synthesized in an objective way. Moreover, such annotations may be attached not only to digital documents by annotations systems but also to physical objects such as things, places ... Annotating with hand-held devices such as daily used mobile phones is already feasible (Hansen, 2006).

References

Adler, M. J. and van Doren, C. (1972). *How to Read a Book*. Simon & Shuster, NY.

Agosti, M. and Ferro, N. (2005). Annotations as Context for Searching Documents. In *CoLIS '05: Proceedings of the 5th International Conference on Conceptions of Library and Information Sciences*, volume 3507 of *LNCS*, pages 155–170. Springer.

Agosti, M. and Ferro, N. (2006). Search Strategies for Finding Annotations and Annotated Documents: The FAST Service. In *FQAS '06: Proceedings of the 7th International Conference on Flexible Query Answering Systems*, volume 4027 of *LNCS*, pages 270–281. Springer.

[12] Wikipedia is a multilingual, web-based, free content encyclopedia project cf. `http://en.wikipedia.org`.

Cabanac, G., Chevalier, M., Chrisment, C., and Julien, C. (2005). A Social Validation of Collaborative Annotations on Digital Documents. In *International Workshop on Annotation for Collaboration*, pages 31–40, Paris. Programme société de l'information, CNRS.

Cabanac, G., Chevalier, M., Chrisment, C., and Julien, C. (2006). Validation sociale d'annotations collectives : argumentation bipolaire graduelle pour la théorie sociale de l'information. In *INFORSID'06 : 24e congrès de l'INFormatique des Organisations et Systèmes d'Information et de Décision*, pages 467–482. Éditions Inforsid.

Cayrol, C. and Lagasquie-Schiex, M.-C. (2005a). Gradual Valuation for Bipolar Argumentation Frameworks. In Godo, L., editor, *Proceedings of the European Conference on Symbolic and Qualitative Approaches to Reasoning and Uncertainty (ESQUARU)*, volume 3571 of *LNCS*, pages 366–377. Springer.

Cayrol, C. and Lagasquie-Schiex, M.-C. (2005b). Graduality in Argumentation. In *Journal of Artificial Intelligence Research*, volume 23, pages 245–297.

Choueka, Y. (2001). Aviezri Fraenkel's Work in Information Retrieval and Related Areas. *Electr. J. Comb.*, 8(2).

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, (20):37–46.

Denoue, L. and Vignollet, L. (2000). An annotation tool for Web browsers and its applications to information retrieval. In *RIAO 2000: Proceedings of the 6th international conference on Information Retrieval and its Applications*, pages 180–196, Paris, France. Le CID.

DeRose, S., Daniel, R., Grosso, P., Maler, E., Marsh, J., and Walsh, N., editors (2002). *XML Pointer Language (XPointer)*. W3C.

Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in a nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77:321–357.

Fleiss, J. L. (1971). Measuring Nominal Scale Agreement Among Many Raters. *Psychological Bulletin*, 76(5):378–382.

Fleiss, J. L., Levin, B., and Paik, M. C. (2003). The Measurement of Interrater Agreement. In Fleiss, J. L., Levin, B., and Paik, M. C., editors, *Statistical Methods for Rates and Proportions – 3rd edition*, chapter 18, pages 598–626. John Wiley & Sons, Inc.

Fraenkel, A. S. and Klein, S. T. (1999). Information Retrieval from Annotated Texts. *J. Am. Soc. Inf. Sci.*, 50(10):845–854.

Frommholz, I. (2005). Applying the Annotation View on Messages for Discussion Search. In Voorhees and Buckland (2005).

Frommholz, I., Brocks, H., Thiel, U., Neuhold, E. J., Iannone, L., Semeraro, G., Berardi, M., and Ceci, M. (2003). Document-Centered Collaboration for Scholars in the Humanities–The COLLATE System. In Koch, T. and Sølvberg, I., editors, *ECDL*, volume 2769 of *Lecture Notes in Computer Science*, pages 434–445. Springer.

Frommholz, I. and Fuhr, N. (2006). Probabilistic, Object-oriented Logics for Annotation-based Retrieval in Digital Libraries. In *JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pages 55–64, New York, NY, USA. ACM Press.

Frommholz, I., Thiel, U., and Kamps, T. (2004). Annotation-based Document Retrieval with Four-Valued Probabilistic Datalog. In *WIRD '04: Proceedings of the first SIGIR Workshop on the Integration of Information Retrieval and Databases*, pages 31–38, Sheffield, UK.

Goguen, J. A. (1997). Towards a Social, Ethical Theory of Information. In Bowker, G., Gasser, L., Star, S. L., and Turner, W., editors, *Social Science Research, Technical Systems and Cooperative Work: Beyond the Great Divide*, pages 27–56. Erlbaum.

Hansen, F. A. (2006). Ubiquitous annotation systems: technologies and challenges. In *HYPERTEXT '06: Proceedings of the 17th conference on Hypertext and hypermedia*, pages 121–132, New York, NY, USA. ACM Press.

Jackson, H. J. (2002). *Marginalia: Readers Writing in Books*. Yale University Press.

Kahan, J., Koivunen, M.-R., Prud'Hommeaux, E., and Swick, R. R. (2002). Annotea: an open RDF infrastructure for shared Web annotations. *Comp. Netw.*, 32(5):589–608.

Karacapilidis, N. and Papadias, D. (2001). Computer supported argumentation and collaborative decision making : the HERMES system. *Information systems*, 26(4):259–277.

Kleiner, I. (2000). From Fermat to Wiles: Fermat's Last Theorem Becomes a Theorem. *Elemente der Mathematik*, 55(1):19–37.

Krippendorff, K. (1980). *Content Analysis: an Introduction to Its Methodology*. Sage Publications, Thousand Oaks, CA.

Lee, J. H. (1997). Analyses of Multiple Evidence Combination. In *SIGIR '97: Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 267–276, New York, NY, USA. ACM Press.

Marshall, C. C. (1998). Toward an ecology of hypertext annotation. In *HYPERTEXT '98: Proceedings of the 9th ACM conference on Hypertext and hypermedia: links, objects, time and space—structure in hypermedia systems*, pages 40–49, New York, NY, USA. ACM Press.

Murphy, P. K., Long, J. F., Holleran, T. A., and Esterly, E. (2003). Persuasion online or on paper: a new take on an old issue. *Learning and Instruction*, 13(5):511–532.

O'Hara, K. and Sellen, A. (1997). A Comparison of Reading Paper and On-Line Documents. In *CHI '97: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 335–342, New York, NY, USA. ACM Press.

O'Reilly, T. (2005). What Is Web 2.0 – Design Patterns and Businness Models for the Next Generation of Software. Available online `http://tim.oreilly.com/lpt/a/6228`.

Ovsiannikov, I. A., Arbib, M. A., and McNeill, T. H. (1999). Annotation technology. *Int. J. Hum.-Comput. Stud.*, 50(4):329–362.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: Sentiment Classification using Machine Learning Techniques. In *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, pages 79–86, Morristown, NJ, USA. Association for Computational Linguistics.

Röscheisen, M., Mogensen, C., and Winograd, T. (1994). Shared web annotations as a platform for third-party value-added, information providers: Architecture, protocols, and usage examples. Technical report CSDTR/DLTR, Stanford, CA, USA.

Sellen, A. J. and Harper, R. H. (2003). *The Myth of the Paperless Office*. MIT Press, Cambridge, USA.

Thiel, U., Brocks, H., Frommholz, I., Dirsch-Weigand, A., Keiper, J., Stein, A., and Neuhold, E. J. (2004). COLLATE – A collaboratory supporting research on historic European films. *Int. J. Digit. Libr.*, 4(1):8–12.

Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E., and Ciravegna, F. (2006). Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Web Semantics: Science, Services and Agents on the World Wide Web*, 4(1):14–28.

Vasudevan, V. and Palmer, M. (1999). On Web Annotations: Promises and Pitfalls of Current Web Infrastructure. In *HICSS '99: Proceedings of the 32nd Annual Hawaii International Conference on System Sciences*, volume 2, page 2012 (9 pages), Washington, DC, USA. IEEE Computer Society.

Voorhees, E. M. and Buckland, L. P., editors (2005). *The Fourteenth Text REtrieval Conference (TREC 2005)*, Gaithersburg, MD, USA. NIST.

Wolfe, J. (2002). Annotation technologies: A software and research review. *Computers and Composition*, 19(4):471–497.

Wolfe, J. L. and Neuwirth, C. M. (2001). From the Margins to the Center: The Future of Annotation. *Journal of Business and Technical Communication*, 15(3):333–371.