Towards Best Practice Standards for Enhanced Knowledge Discovery Systems

Roland Stuckardt

Johann Wolfgang Goethe University Frankfurt am Main
D-60433 Frankfurt am Main, Germany
roland@stuckardt.de

Abstract

Assessing enhanced knowledge discovery systems (eKDSs) constitutes an intricate issue that is understood merely to a certain extent by now. Based upon an analysis of why it is difficult to formally evaluate eKDSs, it is argued for a change of perspective: eKDSs should be understood as intelligent tools for qualitative analysis that support, rather than substitute, the user in the exploration of the data; a qualitative gap will be identified as the main reason why the evaluation of enhanced knowledge discovery systems is difficult. In order to deal with this problem, the construction of a *best practice model* for eKDSs is advocated. Based on a brief recapitulation of similar work on spoken language dialogue systems, first steps towards achieving this goal are performed, and directions of future research are outlined.

1. Elaboration of Problem Statement

The user-oriented assessment of enhanced knowledge discovery systems is a sophisticated problem that is understood merely to a certain extent by now. It imposes a series of *challenges* for which no ready-made solutions are available:

1. In contrast to less complex applications, there is no direct correlation between the performance of the natural language processing base technology and the usability as perceived by the user. For applications such as spell checking and voice recognition, quantitative evaluation measures (percentage of recognized incorrectly spelled words, transcription accuracy) can be expected to correlate quite well with perceived usability. In contrast, for enhanced knowledge discovery systems, no suitable quantitative criteria seem to be readily available.

2. It is difficult to formally define a prototypical task that matches the knowledge discovery needs of all, or at least of a large fraction of users. Too much depends on the specific application scenarios (and of their user-specific perception), which seem to be difficult to standardize and, hence, to model beforehand.

3. Enhanced knowledge discovery typically works on huge amounts of data. Due to this *and to the complexity of the knowledge discovery task*, it is regarded to be unfeasible to construct respective reference data intellectually through human annotators. In this regard, it is important to understand the difference to restricted knowledge discovery scenarios such as basic information extraction, the task of which consists in the document-*local* combination of information only, which may, with considerable efforts, be modeled by suitable text annotation schemes. This is impossible with enhanced knowledge discovery, which, in general, involves relating information contributed by different documents.

4. The homogenity of the data may vary, particularly regarding *type* (e.g., domain and genre of documents) and *reliability*. In contrast to the prototypical application cases that have been considered during the classical evaluation studies (such as the Message Understanding Conferences (MUCs), cf. (MUC 7, 1998; MUC 6, 1996)), the document sets to be processed are not necessarily well-behaved. In particular, they may contain non-factual texts that express differing opinions or points of view on a particular topic. This makes the task of constructing reference data considerably harder.

5. The data as well as its homogenity may vary over time, as in the case of web-based knowledge discovery applications. The same may hold with respect to the typical tasks of the users.

From all this, it follows that it is difficult to define how a "good" output of the knowledge discovery process looks like. Tasks like the identification of market trends seem to be simply too abstract to arrive at a level of formal transparency as achievable for more restricted tasks.
The subsequent sections elaborate upon the issues pointed out above. In section 2., previous work on formal evaluation is recapitulated; in particular, the notions of intrinsic vs. extrinsic evaluation are discussed and related to the problem of assessing enhanced knowledge discovery systems. Building up on this analysis, section 3. proposes a change of perspective: enhanced knowledge discovery applications should be considered as tools that, in large parts, assist in rather than carry out for themselves the analysis of the data: *as enhanced browsers for the qualitative exploration of data, they support rather than substitute the user*, who hence remains responsible for the central intellectual component of the task. This leads to the identification of the *qualitative gap* (section 4.), which will be singled out as the main reason why the evaluation of enhanced knowledge discovery system is difficult. Section 5. draws some important conclusions and suggests promising ways to deal with this problem. In particular, it is argued in favour of the statement of *best practice guidelines* for enhanced knowledge discovery applications. Based on a brief recapitulation

of similar work on spoken language dialogue systems, first steps towards achieving this goal are performed, and directions of future research are outlined.

2. Intrinsic vs. Extrinsic Evaluation

According to, e.g., (Mani, 2002), efforts of evaluating natural language processing systems may be categorized along various dimensions. The intrinsic vs. extrinsic distinction turns out to be of particular importance here:[1]

> **Intrinsic** *evaluations test the system in itself;* **extrinsic** *evaluations test the system in relation to some other task [...].*

According to the above problem statement, tasks to be assisted by the application of enhanced knowledge discovery systems are typically too complex in order to infer application performance from experiments at intrinsic evaluation level only. On the other hand, *generic* extrinsic evaluation imposes problems as well, since, as initially identified, standardizing the knowledge discovery task across users and across specific application scenarios is regarded to be unfeasible in many cases. Clearly, it is possible to extrinsically evaluate systems in *specific* application contexts. However, results are unlikely to generalize; thus, such evaluations cannot be taken as expressive substitutes of in-situ experiments in the particular application scenario an enhanced knowledge discovery system is aimed for.

So how to deal with this situation, according to which, in the case of *enhanced* knowledge discovery tasks, intrinsic evaluation is feasible, but not sufficiently expressive, whereas extrinsic evaluation would be expressive, but is not expected to yield results that generalize across users and application scenarios? Let's take a closer look at the issue why extrinsic evaluation is unlikely to yield once-for-all predictions regarding the performance of enhanced knowledge discovery systems. Figure 1 illustrates the generic application scenario of knowledge discovery systems. The input to the system consists of potentially heterogeneous collections of source documents, which might contain texts as well as tabular data and graphics. These documents are submitted to the knowledge discovery application system, which, possibly driven by a user query, yields an output that can be considered as a *view* on the input document collection. There are many types of operations that might be involved to generate this view, be it textual or graphical information extraction, information retrieval, data mining, or categorization based on similarity criteria.

The essential distinction, however, regards two different stages of processing: (1) the *algorithmic symbol transformation* performed by the knowledge discovery system, which comprises the different types of operations mentioned before, (2) and the *qualitative intellectual interaction* of the understanding user with the system, which comprises the statement of appropriate queries, the analysis of the output, eventually followed by the drawing of conclusions regarding the particular knowledge discovery need. By definition, intrinsic evaluation is related to the processing at the algorithmic stage. In general, these evaluation experiments are based on intellectually annotated corpora and

[1](Mani, 2002), p. 223-4, typographical emphasis by Mani

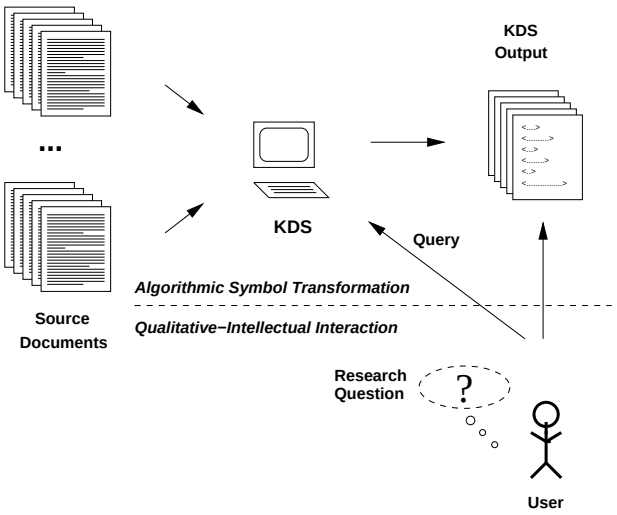

Figure 1: Generic application scenario of KDSs

on formally defined performance measures, which can be computed without further human involvement. In contrast, extrinsic evaluations refer to the output at the qualitative-intellectual stage of analysis, which might aim at the solution of quite abstract and heterogeneous tasks. This leads to a central observation regarding why evaluation of enhanced knowledge discovery systems is hard:

> *Enhanced knowledge discovery systems typically cannot be evaluated extrinsically according to general standards because the main surplus value of the knowledge discovery process is generated in a heterogeneous way at the qualitative-intellectual stage of analysis.*

3. Change of Perspective: Enhanced KDSs Support Qualitative Analysis

This hints at adopting a different perspective: enhanced knowledge discovery systems should *not* be looked at in the same way as at their ancestors with restricted scope, e.g. textual information extraction systems as considered during the MUCs, which can be meaningfully assessed by intrinsic evaluation. Instead, they should be understood as intelligent browsers that support, rather than substitute, the user in the qualitative exploration of the data.

In this sense, their contribution is similar to the contribution of software solutions for computer-*supported* content analysis, which are employed in the Social Sciences (e.g. (Fielding and Lee, 1991; Huber, 1992)). Essentially, these systems assist the user in retrieving and browsing data that might be relevant with respect to the specific research question. In particular, they provide enhanced functionality for the creative-explorative play with the data, such as cut-and-paste tools to manually extract parts of the data and facilities for the intellectual classification of the data according to user-defined categorization schemes. This enables the user to intellectually generate views over the data in order to gain insight in her field of research. This contrasts with computer-*based* content analysis systems, which employ (usually elementary) *automatic* categorization of textual data (cf. (Mohler, 1989)): while software systems that

support content analysis might offer tools for retrieving relevant content as well,[2] the central decisions of how to classify the data and of which conclusions to draw regarding the research question are left to the discretion of the user.

4. The Qualitative Gap

Thus, as in the case of software systems for computer-*supported* content analysis, rather than aiming at an automatic deep analysis the output of which is near to the answer, enhanced knowledge discovery systems are designed to foster intellectual understanding. Regardless of whether one subscribes to the point of view that there is a principal upper bound concerning the algorithmic explicability of cognitive processes, this can be interpreted as acknowledgement of the fact that the scope of algorithmic knowledge discovery will always be limited due to restricted coverage of state-of-the-art technology as well as due to practical feasibility issues, and that the actual *understanding* of the data remains up to the discretion of the user anyway. Based on these observations, the notion of the *qualitative gap* can be defined:

> *Software solutions that support the intellectual exploration of the data through the user, such as enhanced knowledge discovery systems, implicitly acknowledge the existence of a qualitative gap, which is due to practical limitations of the technology for automatically identifying relevant content: to optimally support the user in gaining insight in particular fields of research, tools for automatic content analysis are supplemented with features that allow for creatively browsing the data. Since, typically, the qualitative gap between the scope of algorithmic analysis and the requirements according to the research topics to be investigated is considerable, intrinsic evaluation can be expected to be not expressive enough.*

As further argued in section 2., whether generic extrinsic evaluation could be employed instead depends upon whether the knowledge discovery task (in particular, its qualitative component) can be standardized across users and across specific application scenarios.

5. Implications - Towards Best Practice Guidelines for Developing eKDSs

The above discussion has revealed that one central property of enhanced knowledge discovery systems is the typically considerable gap between the contributions of the individual technology components and the (typically quite abstract) insight into the research topic gained by the user through usage of the system as a tool that supports, rather than substitutes, understanding of the data. Regarding the issue that the components that can be subjected to intrinsic evaluation contribute only quite indirectly to the success in particular application scenarios, enhanced KDSs closely resemble other classes of complex Natural Language Technology applications.

[2] be it basic string search or enhanced retrieval and extraction functionality

5.1. Best practice guidelines for NLP applications: objectives and development issues

A related topic has been investigated for *spoken language dialogue systems* (SLDS), which exhibit the analogous property that their usability and perceived degree of usefulness highly depends on the particular application context, i.e. on the information needs and on the communicative or interactional preferences of the typical user. This identification of a gap between technology-related intrinsic criteria and the observed usefulness at extrinsic level has led to the development of *best practice guidelines* for spoken language dialogue systems, which, according to (van Kuppevelt et al., 2000), p. 207, are to be understood as

> *[...] a mapping from functional parameters to parameters of design and development*

Developing best practice guidelines hence means (ibd., p. 207f),

> *[...] to determine exactly what the mapping is like, and how its salient properties are best explained to a broad spectrum of laymen and professionals who find themselves confronted with the problem of getting an SLDS that answers their needs.*

According to these definitions, best practice guidelines have general scope in the sense that they are intended to fulfil the requirements of all involved stakeholders, viz. deployers, developers, customers, and users, i.e. (ibd, p. 206)

> *[...] to enable them to make accurate and successful design and implementation decisions, in accordance with broad consensus of what must be best practice in this particular engineering domain.*

Acknowledging the intricacies of designing appropriate spoken language dialogue interfaces thus in effect amounts to recognizing this issue as a creative intellectual engineering activity based on well-founded standards rather than as a matter of solid craftsmanship that merely relies upon the application of basic schematic knowledge.

The above discussion urges upon the conclusion that the statement of best practice guidelines should be the approach of choice for coping with the challenges of development, selection, and optimization of enhanced knowledge discovery systems. As in the case of spoken language dialogue systems, the degree of success of a solution highly depends on factors determined by the particular application context. Hence, the postulated objectives for the development of best practice guidelines for SLDSs can be taken as guiding principles for respective work on eKDSs. According to (van Kuppevelt et al., 2000), best practice guidelines should cover three closely related issues: (1) *stock-taking of the state-of-the-art* in order to enable the stakeholders to quickly inform themselves about the range of options for design, implementation, and evaluation; (2) *quality control* through the provision of criteria that support the selection of options that are best suited to particular application requirements; (3) *economic control*, to be achieved by making available a repository of resources in order to foster the

reuse of existing components, design know-how, and gathered experience.
Hence, as required for dealing with the typical scenario sketched in the workshop description, best practice models in particular provide criteria for the design of optimal solutions that fit best within particular application contexts.

5.2. Best practice guidelines for eKDSs

Thus, instead of entering into the ad-hoc discussion of how to deal with this typical scenario, a principled approach is advocated. The DISC best practice model, which covers the three above issues identified by (van Kuppevelt et al., 2000), is centered around a series of fundamental *aspects* of SLDS (system components as well as abstract development issues); it discusse them along a common scheme of *main items* (cf. (DISC, 2000)).[3]
It is proposed to take this approach as the point of departure for respective work on eKDSs. Regarding enhanced knowledge discovery systems, some main *aspects* are:

1. *Information Extraction Engines*, qualified by type of data (textual, graphical etc.),
2. *Information Retrieval Engines*, qualified by type of data,
3. *Data Mining Engines*, qualified by type of data,
4. *Categorization Engines*, qualified by type of data,
5. *Indexing Schemes*, qualified by type of data
6. *Query Engines*, qualified by type of data
7. *Knowledge Sources*, e.g., supported types of data, covered encoding schemes (.doc, .pdf, .ps, email archives, .jpeg, .tiff, etc.), static vs. dynamic data, intranet and/or internet resources etc.), amount of data to be processed, reliability and homogenity issues,
8. *Graphical User Interface*,
9. *Human Factors* (types of users, their degree of experience and previous knowledge etc.),
10. *Systems Integration*,
11. *Knowledge Discovery Objective* (as specific as possible, as abstract as necessary).

While some of the abstract aspects are immediately adopted from the SLDS best practice model (*Human Factors*, *Systems Integration*, the more concrete ones are not, as they correspond to specific system components of knowledge discovery systems without counterpart in the realm of dialogue systems. There is a further important difference that should be noticed here: regarding eKDSs, the extent to which particular systems differ with respect to the individually relevant aspects is considerably larger than regarding SLDS, which typically instantiate *all* aspects of their best practice model. A particular knowledge discovery solution might include an information extraction engine for graphical data, whereas another system might cover *textual* input only. Hence, the recommendations provided for the *Systems Integration* aspect are necessarily situated at a more abstract level; they strongly interdepend with the particular knowledge discovery objective, which, as a consequence, should be covered by a separate dedicated aspect. Again, this illustrates that, compared to many other natural language engineering problems, the development of eKDSs is a particularly intricate matter.
As far as applicable, each aspect should be discussed along several dimensions[4]: (a) grid (factual properties), (b) life cycle (development issues), (c) evaluation, (d) checklists, (e) glossary, and (f) references. Much specific previous work has been done on these issues. For instance, it might be referred to the experiences and resources gathered at the various DARPA- and EC-funded evaluation contests, e.g. TREC (information retrieval) and MUC (information extraction). Thus, to a large extent, modeling best practice amounts to an in-depth analysis of the state-of-the-art of the above-identified aspects of knowledge discovery solutions. Further aligning these different sources of knowledge according to the standardized scheme of a best practice model necessitates a considerable research effort.

6. The Next Steps

Proceeding along similar lines as followed during development of the DISC model, the elaboration of the best practice guidelines for eKDSs might be accomplished in three stages: (a) analysis of the state-of-the-art through data collection from different evaluation sources, (b) identification of particular constraint-oriented (i.e. application context sensitive) evaluation criteria, and (c) criteria integration, the output of which consists in the best practice methodology proper that provides high-level criteria for the informed choice among the technological options. Obviously, the last-mentioned stage embodies the major intellectual challenge.
According to the above considerations, modeling best practice regarding eKDS can be regarded to impose even more intricate challenges than in the case of SLDSs. Mainly due to the qualitative gap, the extent to which particular eKDSs differ with respect to their individual relevant aspects is considerably larger. Thus, the development of a sufficiently expressive best practice model constitutes a major research effort that should be addressed by a joint project with partners from commercial as well as non-commercial research, comprising all types of stakeholders (developers, deployers, customers, users). This project is necessarily interdisciplinary, as it covers issues from a broad range of disciplines (linguistic and mathematical models of content analysis, software system engineering, human-computer interaction).

[3] In accord with its objective, the resource repository of the DISC best practice guide has been made freely available at the web page (DISC, 2000). DISC is extensively documented in numerous online and offline publications, e.g. the deliverables made available at (DISC, 2000) or the book (Bernsen et al., 1998).

[4] according to the DISC terminology, *main items*

7. References

Bernsen, Nils Ole, H. Dybkjaer, and Laila Dybkjaer, 1998. *Designing Interactive Speech Systems: From First Ideas to User Testing*. Berlin, Heidelberg: Springer Verlag.

DISC, 2000. The disc best practice guide. available (March 1st, 2004) at http://www.disc2.dk/.

Fielding, Nigel G. and Raymond M. Lee, 1991. *Using Computers in Qualitative Research*. SAGE Publications.

Huber, Günter L., 1992. *Qualitative Analyse. Computereinsatz in der Sozialforschung.*. München, Wien: R. Oldenbourg Verlag.

Mani, Inderjeet, 2002. *Automatic Summarization*. Amsterdam/Philadelphia: John Benjamins.

Mohler, Peter Ph., 1989. Computergestützte inhaltsanalyse: überblick über die linguistischen leistungen. In Batori and Lenders (eds.), *HSK 4. Handbuch zur Sprach- und Kommunikationswissenschaft*.

MUC 6, 1996. *Proceedings of the Sixth Message Understanding Conference (MUC-6)*. Morgan Kaufmann.

MUC 7, 1998. *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. Published online, formerly (December 9, 1999) available at http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/co_task.html.

van Kuppevelt, Jan, Ulrich Heid, and Hans Kamp, 2000. Best practice in spoken language dialogue systems engineering - introduction to the special issue. *Natural Language Engineering*, 6(3 & 4):205–212.