Mining Documents for Complex Semantic Relations by the Use of Context Classification

Andreas Schmidt and Markus Junker

German Research Center for Artificial Intelligence (DFKI)
P.O. Box 2080, 67608 Kaiserslautern / Germany
{schmidt,junker}@dfki.uni-kl.de

Abstract. Causal relations symbolize one of the most important document organization and knowledge representation principles. Consequently, the identification of cause-effect chains for later evaluation represents a valuable document analysis task. This work introduces a prototype implementation of a causal relation management and evaluation system which functions as a framework for mining documents for causal relations. The central part describes a new approach of classifying passages of documents as relevant considering the causal relations under inspection. The "Context Classification by Distance-Weighted Relevance Feedback" method combines passage retrieval and relevance feedback techniques and extends both of them with regard to the local contextual nature of causal relations. A wide range of parameter settings is evaluated in various experiments and the results are discussed on the basis of recall-precision figures. It is shown that the trained context classifier represents a good means for identifying relevant passages not only for already seen causal relations but also for new ones.

1 Introduction

Mining documents for causal relations [1] represents a worthwile document analysis task and can be regarded as a predecessor step towards "knowledge-based document analysis and understanding" [2] and semantic net constructions for a better understanding and superior view of the context of document collections.

Figure 1 shows a semantic net which reflects cause-effect chains that are typical for applications in the economic science domain, i.e. the so-called "Scenario-Management" [3]. The goal of this paper is to automatically identify passages that contain attributes such as "domestic demand" and "interest rate" as causes and / or effects of a specific causal relation, i.e. "influence: interest rate influences domestic demand" and to incrementally learn a classifier by relevance feedback. The application of this algorithm helps to mine new relations from text collections and makes it possible to build up a causal semantic net.

The subsequent work consists of four sections. First of all, the "System Design of the Causal Relation Management and Evaluation System" (CRMES) section describes the main steps of the causal relation mining process and the respective CRMES application. Then, the "Context Classification by Distance-Weighted Relevance Feedback" chapter details the theoretic background of the new developed algorithm.

D. Lopresti, J. Hu, and R. Kashi (Eds.): DAS 2002, LNCS 2423, pp. 400–411, 2002.

The forth chapter describes the experimental design followed by a discussion of the experimental results. The last chapter concludes the paper with a summary of the results and an outlook for future research.

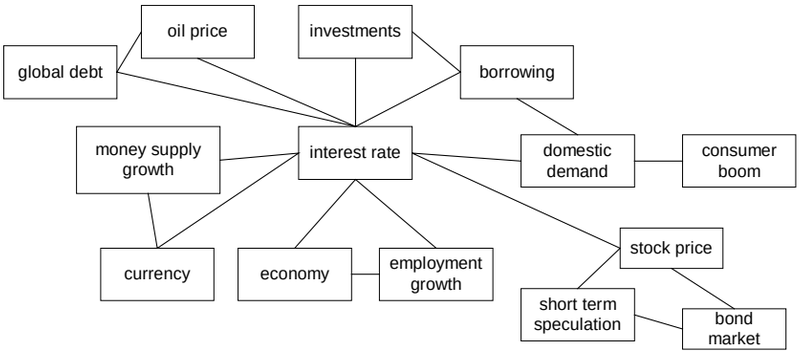

Fig. 1. Causal Relations visualized as a Semantic Net

2 System Design of the Causal-Relation-Management and Evaluation System

This chapter describes the overall process and application scenario for the "Causal-Relation Management and Evaluation System" (CRMES)[1]. Figure 2 gives an overview of the passage mining process steps. Figure 3 depicts a screenshot of the CRMES. Input to the CRMES is a base text collection. Several retrieval models - here a boolean retrieval model and a passage vector-space model with and without relevance feedback, and a parameter base for full parametrization control the system behavior. Furthermore a set of distance-weighted context relevance feedback classifiers and a set of causal relation attributes consisting of cause- and effect-terms influence the specific contextual causal relation setting.

Given one is interested in a causal relation with the cause being "interest rate" and the effect not further defined (A) q_{CE}:[CAUSE="interest rate" $\Rightarrow$ EFFECT=""]. With respect to the retrieval model (B), the parameter base (C) and the current distance-weighted relevance feedback context classifier (D), CRMES returns a ranked list of passages that may contain relevant causal relations. Passage by passage (E) can be browsed through (F) and may be manually judged relevant or irrelevant by the user (G). The resulting relevant or irrelevant terms of the passage are listed under "Rel. Passage" or "Irrel. Passage" respectively. These judgements are input to the modification of the context relevance feedback classifier as described below.

[1] CRMES is part of ongoing research in the "Adaptive READ" project of the German Research Center for Artificial Intelligence (DFKI): http://www.adaptive-read.de

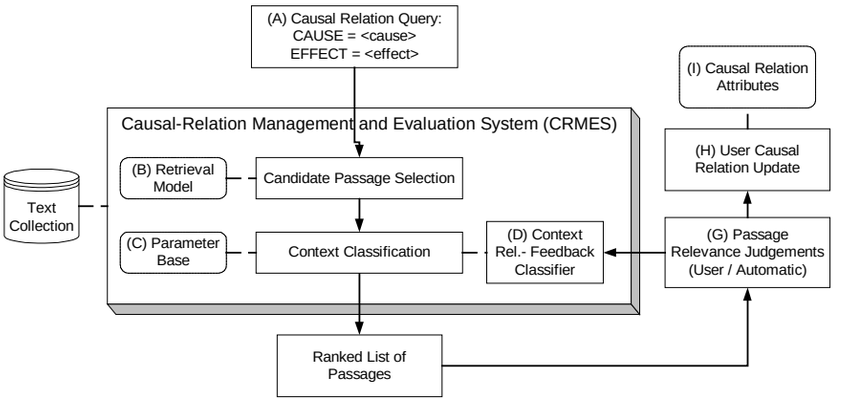

Fig. 2. Passage Mining Process Steps

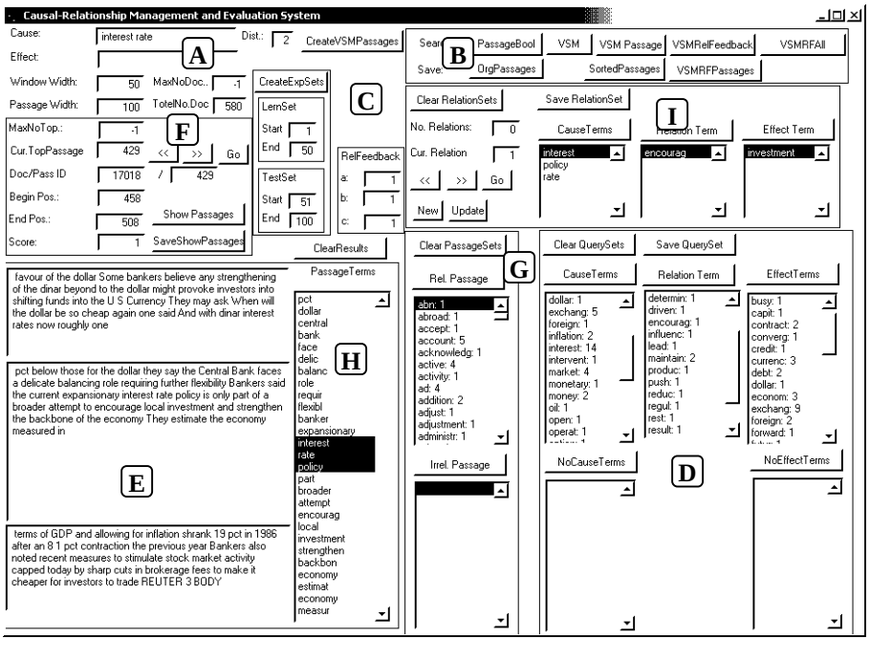

Fig. 3. Causal-Relation Management and Evaluation System (CRMES) Prototype

Furthermore the user can manually update the causal relation attribute set by highlighting relevant passage terms (H) and store them for future reference (I). This way, CRMES not only learns to fine-tune its relevance feedback context classifier with each new relevance judgement, but also extends its knowledge about new relations. The system can also be run in batch mode by using predetermined relevant

and irrelevant passages and specific experimental parameter settings. The batch mode was actually used for the following experiments.

3 Context Classification by Distance-Weighted Relevance Feedback

The new developed "Context Classification by Distance Weighted Relevance Feedback" combines Rocchio relevance feedback methods (see [4], pp. 118) with passage retrieval approaches [5]. First of all the conventional "Document Vector Space Model" [6] as a representation of texts is applied on passage level resulting in a new derivate "Passage Vector Space Model". Second the new Rocchio relevance feedback method incorporates the "distance-weighted" idea in such a way that terms occurring close to query terms are weighted more and as such are more important candidates for query term expansion than those that stand further apart.

Additional relevance feedback and query expansion techniques that base on probabilistic text models can be found in [7]. Other applications of term selection for automatic query expansion are described in [8]. While the term expansion technique in this paper base on learning a classifier, [9] shows interesting approaches in the fields of reinforcement learning. Last but not least, [10] gives a detailed survey of a complete different view how to extract information from texts.

3.1 Passage Vector Space Model

A passage vector space model PVSM consists of a triple

$$PVSM=(P, Q, R(p_j, q_i)) \tag{1}$$

where $P = \{p_1, ..., p_M\}$ is a set of M passages – representing the text-collection, $Q = \{q_1, ..., q_k\}$ is a set of k queries and R is a ranking function which defines an ordering among passages with regard to queries. Let $T = \{t_1, ..., t_N\}$ be the set of all index terms that occur in P where P may be preprocessed to eliminate stopwords (such as "the", "a", etc.).

On the basis of term-occurrence counts, a passage-term frequency matrix PTF: (T x P) is constructed where $PTF(t_i, p_j)$ is the raw frequency of term t_i within passage p_j. The passage-term frequency matrix PTF may be dampened by a dampening function f to reflect the fact that more occurrences of a term indicate higher importance but not as much as the raw frequency might suggest. Typical dampening functions are square root: dPTF=f(PTF) = sqrt(PTF) or log: dPTF=f(PTF)=1+log(PTF). The second frequency count captures the importance of terms across the whole collection. The inverse passage term frequency is defined as

$$\text{iptf}_i = \log \frac{M}{\text{pf}_i} \tag{2}$$

where M = |P| number of passages in the collection and pf_i is the number of passages that term t_i occurs in. IPTF gives full weight for terms that occur in ONE passage and zero weight for those that occur in ALL passages.

The definition of weights $w=w(t_i, p_j)$ which are associated to the index terms t_i and passages p_j combine various combinations of the term-occurrence and passage frequency counts into several PTF.IPTF weighting schemes (similar to the ones based on documents in [4] and [6]). This work uses the subsequent weighting scheme

$$w_{i,k} = dptf_{i,k} \times iptf_i \times \frac{1}{|\vec{w}_k|} \quad . \tag{3}$$

Now each passage p_j and query q_l can be represented by a passage-weight vector.

$$\vec{p}_j = (w_{1,j}, ..., w_{n,j}) \text{ and } \vec{q}_l = (w_{1,l}, ..., w_{n,l}) \tag{4}$$

Finally, a ranking function R needs to be defined which expresses the similarity $sim(p_j, q_l)$ between passages and queries. We have chosen the normalized COSINE-similarity[2]

$$R : sim(\vec{p}_j, \vec{q}_l) = \frac{\vec{p}_j \bullet \vec{q}_l}{|\vec{p}_j| \times |\vec{q}_l|} = \frac{\sum_{i=1}^{n} w_{i,j} \times w_{i,l}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{n} w_{i,l}^2}} \, . \tag{5}$$

3.2 Rocchio-Relevance Feedback as a Distance-Weighted Context Classification Method

The conventional Rocchio-Relevance Feedback utilizes all terms of manual or automatic judged relevant / irrelevant passages $P_{Train,rel}$ / $P_{Train,irrel}$ to expand an original query q_0 into an extended query q_{RF}

$$\vec{q}_{RF} = \alpha \vec{q}_0 + \frac{\beta}{|P_{Train,rel}|} \sum_{p_j \in P_{Train,rel}} \vec{p}_{dist,j} - \frac{\gamma}{|P_{Train,irrel}|} \sum_{p_j \in P_{Train,irrel}} \vec{p}_{dist,j} \tag{6}$$

where $\vec{p}_{dist,j}$ equals the original non-distance-weighted passage vector $\vec{p}_j$ and the Rocchio multipliers α, β, γ parametrize the feedback strategy with the following special cases:

- no feedback: $\alpha>0$ $\beta=0$ $\gamma=0$
- pure positive feedback: $\alpha=0$ $\beta>0$ $\gamma=0$
- pure negative feedback: $\alpha=0$ $\beta=0$ $\gamma>0$

The new distance-weighted relevance feedback computation is based on the assumption that terms t_i which are closer to cause-effect terms q_0 characterize the causal relation in a better way than terms that are further apart.

Figure 4 depicts a symbolic causal relation context where the x-axis represent the position of terms and the y-axis assigns a distance-weighted weight to each term t_i according to a given window-weight function win_fun ∈ {"rectangular", "triangular", "hanning"} and relative to a queried causal relation $q_0 = (q_{0,1}, q_{0,2}, \ldots)$. The window functions are placed at each position of all $q_{0,j}$ with their normalized maximum amounting to one so that the resulting distance-weighted passage vector amounts to

[2] For other similarity measures like DICE coefficient etc. see [4]

$$\vec{p}_{dist} = \left(w_{1,p}, w_{2,p}, ..., w_{n,p}\right) \tag{7}$$

$$w_{i,p} = \sum_{q_{0,j} \in \vec{q}_0} wf(q_{0,j}, t_i); \qquad wf(q,t) = win\_fun(dist(q,t)) \tag{8}$$

with dist(q,t) being the distance between q and t in number of terms.

The three window types allow for different explanations of the impact of terms t_i with respect to the distance they have to q_0. By using the rectangular window it is assumed that all terms within the window width have the same impact on the classification of the causal relation. The triangular window reflects a linear correlation between impact and distance while the hanning window allows closer terms to be of relative more importance than terms that are further apart. The baseline parametrization consists of the rectangular window function with a window-width equal to the passage length.

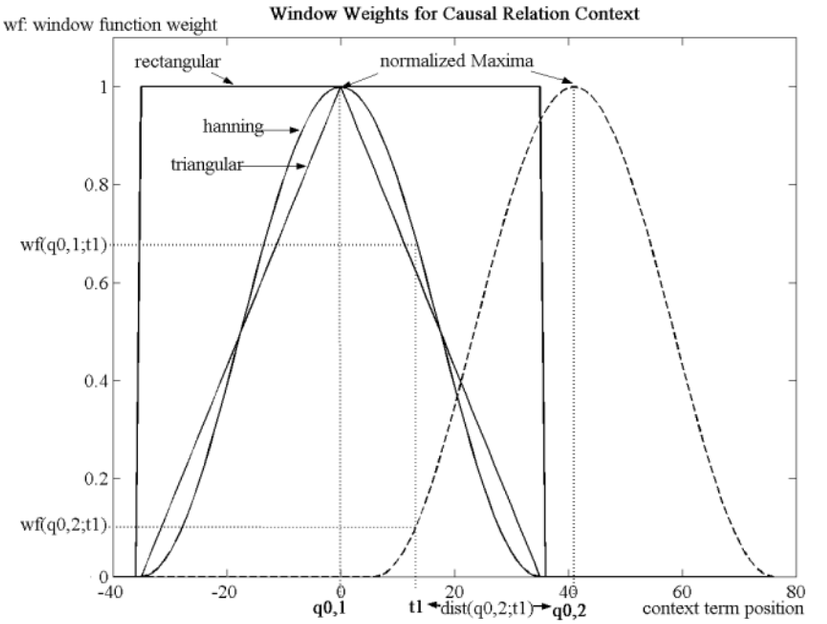

Fig. 4. Distance-Weighted Relevance Feedback Computation

4 Experimental Design

The experimental design is based on running the CRMES system in batch mode where the base text-collection – the REUTERS21578 corpus [11] – consists of 21578 news stories from REUTERS. This collection was transfered into document vector-space representation.

In a first step, all passages that fulfilled a specific causal relation attribute, here: CAUSE = "interest rate" were retrieved and transferred into a passage vector-space model. This procedure translates into posing an initial query q_0 and receiving a ground-truth passageset P_{GT} where each passage is manually judged relevant or irrelevant.

The second step involved the selection of the training- and test-set, P_{Train} and P_{Test}, that were taken aside from P_{GT} for training and testing the distance-weighted context classifier. Altogether four sets were created represented by $P_{Train} = P_{Train,rel} \cup P_{Train,irrel}$ and $P_{Test} = P_{Test,rel} \cup P_{Test,irrel}$ with $P_{Test} \cap P_{Train} = \varnothing$. Table 1 shows an extract of ten causal relation attributes that were part of P_{Train}. Bold rows indicate attributes that also were part of the retrieved testpassages P_{Test} (see table 4 below).

Table 1. Extract of Causal Relation Attributes in P_{Train}

No.	$Cause_{Train}$	$Connective_{Train}$	$Effect_{Train}$
1	**interest rate**	**sluggish**	**global economy**
2	**interest rate policy**	**encourage**	**investment**
3	**interest rate policy**	**strengthen**	**economy**
4	**interest rate**	**reignite**	**inflation**
5	interest rate	curtails	short term speculation
6	oil price	led	interest rate
7	employment growth	led	interest rate
8	interest rate	improve	expectation
9	interest rate		earnings margin
10	interest rate	lead	consumer boom

Third, the Rocchio relevance feedback vector $\vec{q}_{RF}$ as described in equation 6 was trained with P_{Train} according to a variety of parametrization settings. Three dimensions of parametrization were researched in detail:

1. Window functions: rectangular, triangular, hanning
2. Window widths: 10, 20, 30, 40, 50, 100 terms
3. Rocchio multipliers [α, β, γ]:
   - [1, 0, 0]: baseline, original query only
   - [0, 1, 0]: only relevant passages are considered
   - [0, 1, 1]: relevant and irrelevant passages are considered equal; original query is not considered
   - [0, 1, 0.5]: irrelevant passages are considered less important than relevant passages; the original query is neglected
   - [1, 1, 1]: all equal

To reflect an untrained system, the baseline settings consisted of the parametrization set window-function = rectangular (distance between q_0 and respective terms is not taken into consideration) , window-width=100 terms (equals passage width) and Rocchio multiplier = [1, 0, 0] (unexpanded query only). The combination of all parameter settings (three window-functions, six window-widths, five Rocchio sets) led to the sum of 90 different Rocchio relevance feedback vectors.

Finally during the test-phase, the similarity between each relevance feedback vector q_{RF} and each passage $p_{Test,i} \in P_{Test}$ were calculated and ranked using the ranking function R: $sim(p_{Test,j}, q_{RF})$ resulting in a ranked passage list $RP=(rp_1, rp_2, ...)$ where $rp_i = p_j$ with p_j being i-th ranked. The ranked passage list consisted of passages $rp_{i,rel}$ that were relevant, $RP_{rel}=(rp_{1,rel}, ..., rp_{n,rel})$, and passages $rp_{j,irrel}$ that were irrelevant, $P_{irrel} = (rp_{1,irrel}, ..., rp_{m,irrel})$, so that $RP=RP_{rel} \cup RP_{irrel}$.

In order to compare the results, three scenarios were calculated as follows:

1. Modification of the Window Functions:
   fixed window widths and Rocchio-sets; variable window functions
2. Modification of the Window Width:
   fixed window functions and Rocchio-sets; variable window widths
3. Modification of the Rocchio-Set:
   fixed window functions and window widths; variable Rocchio-sets

5 Experimental Results

The results are visualized by precision-recall graphs with precisions being interpolated at 11 standard recall levels. The n-th level $recall_n$ consists of the share of top n ranked relevant passages $|RP_{n,rel}|$ with regard to all relevant testpassages $|P_{Test,rel}|$. The n-th level $precision_n$ is the fraction of $|RP_{n,rel}|$ considering both relevant and irrelevant top n ranked passages $|RP_n|$. Average precision AveP functions as an aggregated measure for comparing the results of the overall classification / retrieval algorithm.

$$\text{recall}_n = \frac{\left|RP_{n,rel}\right|}{\left|P_{Test,rel}\right|};\ \text{precision}_n = \frac{\left|RP_{n,rel}\right|}{\left|RP_n\right|} \tag{9}$$

$$\text{AveP} = \frac{1}{n}\sum_{i=1}^{n}\text{precision}_n \tag{10}$$

Figure 5 shows the precision-recall graphs that correspond to the best two and the worst two average precision results - including the baseline as a representative for the untrained classifier. The corresponding parameter settings can be found in tables 2 and 3. Let the experiment consist of 50 testpassages, $|P_{Test}| = 50$, where 26 of them are judged relevant, $|P_{Test,rel}| = 26$. As an example for interpreting the precision-recall graphs, let the goal be to retrieve 50% of all relevant passages which translates into a recall of 50% or a $|RP_n| = recall_n / |P_{Test,rel}| = 50\%$ x 26 = 13 relevant passages. The untrained classifier reflected by the baseline precision-recall graph reaches a precision of 48% at 50% recall. The total number of ranked passages which needs to be seen amounts to $|RP| = |RP_n| / precision_n = 13 / 48\% = 27$ passages. The best trained

classifier reflected by the Top 1 precision-recall graph reaches a precision of 86% at a recall level of 50% so that the necessary total number of ranked passages amount to |RP| = 13 / 86% = 15 passages only. Compared to the baseline, the best trained classifier allows for a cut of almost half the size in ranked passages to be seen.

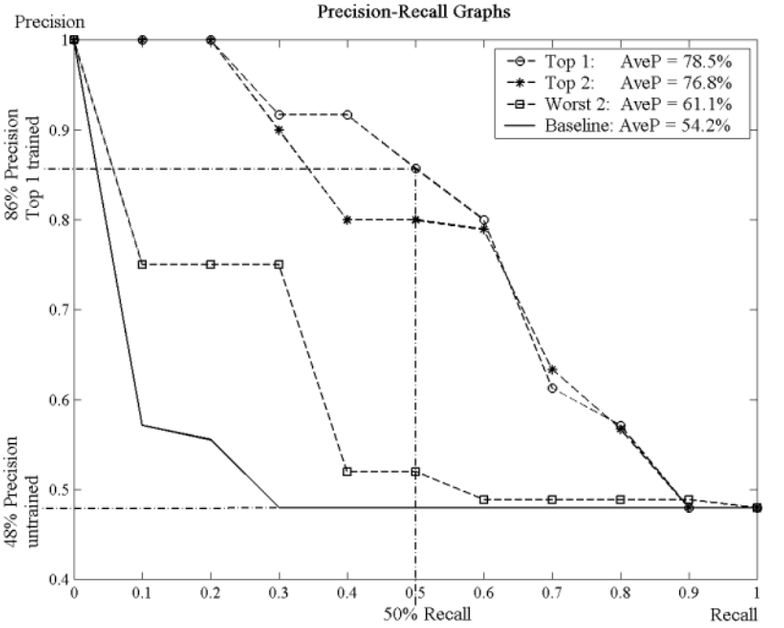

Fig. 5. Best / Worst Precision-Recall Graphs

The parameter settings leading to the best five and the worst five average precisions out of the total 90 parameter settings are listed in tables 2 and 3. The optimization trend seems to be to use a small window-width together with the relevant terms feeded back only (Rocchio-Parameterset = [0, 1, 0]), while the choice of the window function does not seem to have such a big impact.

Table 2. Top five parameter settings

No.	AveP	Window-Type	Window-Width	Rocchio-Parameterset		
				α	β	γ
1	0,785	rectangular	10	0	1	0
2	0,768	hanning	20	0	1	0
3	0,767	triangular	20	0	1	0
4	0,763	rectangular	20	0	1	0
5	0,762	triangular	30	0	1	0

Table 3. Worst five parameter settings

No.	AveP	Window-Type	Window-Width	Rocchio-Parameterset		
				α	β	γ
1	0,542	rectangular	100	1	0	0
2	0,611	rectangular	30	0	1	1
3	0,611	rectangular	30	1	1	1
4	0,623	rectangular	40	0	1	1
5	0,623	rectangular	40	1	1	1

Figure 6 breaks the results of the preceding chapter down on cause-effect level. The "Training Phase" figure section shows the direct cause-effect relations that were part of the passages used to construct the distance-weighted context classifier by considering all fifty training passages. The five bold causal attributes reflect the situation where only ten passages were considered for training (baseline setting).

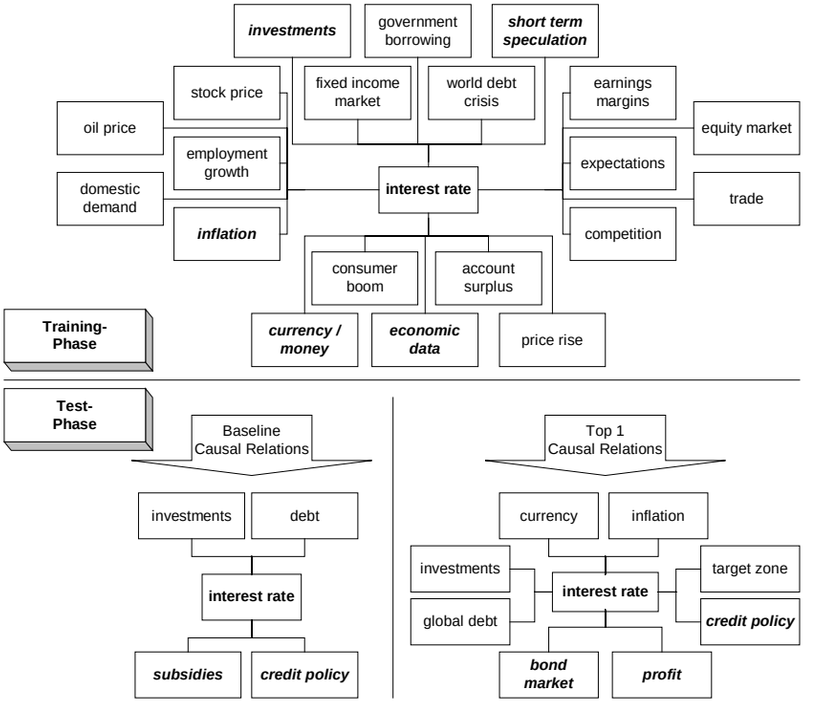

Fig. 6. Comparison of Causal Relations Context Classifier Parametrizations

The "Test Phase" lower part of the figure depicts the cause-effect relations when taking the top ten ranked passages of the baseline context classifier ("No. 1" parameter settings in table 3) and the top 1 context classifier ("No. 1" parameter settings in table 2). Figure 6 confirms the results of the preceding chapter. The optimized parametrization of the top 1 context classifier contains eight cause-effect

relations while the baseline classifier returns only four among the first ten ranked passages.

Table 4 juxtaposes three sample relations that were part of both training and test passages. Passage set one depicts a typical "same-direction" relation with Train: ["**High** interest rate" ⇒ "**sluggish** global economic growth"] and Test: ["interest rate **increases**" ⇒ "**deteriorating** global debt and economic situation"]. Passage set two shows a counter-direction with Train: ["**lower** interest rates" ⇒ "**reignite** inflation"] and Test: ["**rise** in interest rates" ⇒ "**dampening** inflationary pressure"]. Passage set three exemplifies a similarity-based relation where "interest rate policy" in the training passage is set similar to "interest rate" in the test passage.

Table 4. Causal Relations in both Training and Test Passages

	Training Passage	**Test Passage**
1	"… **High interest rates sluggish global economic growth** and creeping protectionism are deepening the **third world debt crisis** …"	"…Amid **new concerns about** inflation, **interest rate increases** and trade, finance ministers and central bankers meet next week to discuss a **deteriorating global debt** and **economic situation**…"
2	"… Bankers said, the government s policy of **fostering lower interest rates could** lead to a consumer boom and **reignite inflation** …"	"…**Rises in interest rates aimed at dampening inflationary pressures** also slow domestic demand …"
3	"… Bankers said, the current **expansionary interest rate policy** is only part of a broader attempt to **encourage local investment** and **strengthen the backbone of the economy** …"	"…Steamship Co Ltd lt ADSA S said, it was **looking to the British market for future investment in view of** high share prices and **interest rates** in Australia…"

However not only already known training phase causal relations are identified but also new ones are mined from the first ten passages ranked by the top 1 context classifier. Table 5 depicts exemplary the three passages that are in bold typeset in figure 6: ["interest rate approach lower level" ⇒ "ease credit policy"], ["lingering anxiety with interest rates" ⇒ "prompt investors to take profits"] and ["weak bond market" ⇒ "rising interest rates"].

Table 5. New mined Causal Relations among Top-Ranked Passages of Top 1 Concept Vector

	Test Passage
1	"… there is little room left for the central bank to further **ease its credit policy as interest rate** levels are now **approaching their lower limit** …"
2	"… Wall Street s **lingering anxiety with interest rates** and inflation **prompting investors to take profits** …"
3	"… A **weak bond market ignited concern about rising interest rates** and inflation …"

6 Conclusion and Outlook

This work presented the "Causal Relation Management and Evaluation System" as a prototype implementation for identifying and managing causal relations in text collections. The new developed method of "Context Classification by Distance-Weighted Relevance Feedback" demonstrated good performance for classifying passages as containing not only already seen relevant causal relations but also unseen examples of causal relations. Consequently, this new approach may be regarded as a worthwile contribution for developing a new kind of knowledge-centered document analysis systems.

Future research may deal with the transitivity of known relations, i.e. "$[a \Rightarrow b \wedge b \Rightarrow c] : \Rightarrow [a \Rightarrow c]$", and the vector-space similarity of terms, i.e. "$[a \Rightarrow b \wedge sim(a,c) \wedge sim(b,d)] :\Rightarrow [c \Rightarrow d]$" to capture a wider spectrum of document understanding and knowledge representation. Another fruitful undertaking may be the application of the new developed context classifier in other research areas that base on the contextual nature of relational elements, such as word-sense disambiguation.

References

1. P.G. Meyer. The relevance of causality. in: *E. Couper-Kuhlen. Cause – Condition-Concession – Contrast: Cognitive and Discourse Perspective*. Mouton de Gruyter, Berlin, New-York, 2000, pp. 9–34
2. C. Wenzel, H. Maus. An Approach to Context-driven Document Analysis and Understanding. in: *4th IAPR International Workshop On Document Analysis Systems – DAS'2000*, Rio de Janeiro, Brazil, Dec. 2000, pp. 121–133
3. J. Gausemeier, A. Fink, O. Schlake. Szenario-Management: Planen und Führen mit Szenarien, 2. Edition, Hanser Verlag München, 1996
4. R. Baeze-Yates, B. Ribeiro-Neto. Modern Information Retrieval. Addison-Wesley, 1999
5. K. Kise et al. Passage-Based Document Retrieval as a Tool for Text Mining with User's Information Needs. in: *K. P. Jantke, A. Shinohara (eds.) Discovery Science, 4th International Conference, DS 2001*, Washington. Lecture Notes in Computer Science, Springer-Verlag, 2001, pp. 155–169
6. C. D. Manning, H. Schütze. Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, 2000
7. Y. Ogawa et al. Structuring and Expanding Queries in the Probabilistic Model. in: *Proceedings of the 9th Text Retrieval Conference (TREC-9)*. NIST Special Publication, 2001
8. M. Adriani, C.J. van Rijsbergen. Informative term selection for automatic query expansion. in: *S. Abiteboul, A.-M. Vercoustre (eds.). ECDL 1999*, Springer-Verlag, Berlin-Heidelberg, pp. 311–322
9. L. Kaelbling, M. Littman. Reinforcement Learning: A Survey. in: *Journal of Artificial Intelligence Research 4 (1996)*, pp. 237–285, Morgan Kaufmann Publishers
10. I. Muslea. Extraction Patterns for Information Extraction Tasks: A Survey. in: *American Association for Artificial Intelligence, 1999*
11. REUTERS Corpus: http://about.reuters.com/researchandstandards/corpus/